VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

PH model v analýze přežití

PH model in survival analysis

2014 Bc. Žaneta Miklová

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Zadání diplomové práce

Student: **Bc. Žaneta Miklová**

Studijní program: N2647 Informační a komunikační technologie

Studijní obor: 1103T031 Výpočetní matematika

Téma: PH model v analýze přežití
PH model in survival analysis

Zásady pro vypracování:

Práce je zaměřena na analýzu přežití v oblasti lékařských dat, kdy přežívání závisí na více proměnných. Pro účely transformace dat o přežití bude použit regresní PH model (Proportional Hazards Model).
Postup práce:

1. Analýza přežití - studium základů.
2. Regresní modely v analýze přežití – přehled modelů.
3. PH model, odhad parametrů v PH modelu.
4. Kvantifikace funkce přežití v PH modelu.
5. Aplikace na lékařská data dodaná vedoucím práce.
6. Testování adekvátnosti modelu.
7. PC implementace, závěry k dodaným datům.

Seznam doporučené odborné literatury:

• Hosmer D.W., Lemeshow S., May S. Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Second Edition. Wiley, Hoboken, NJ, 2008.
• Briš R., Litschmannová M., STATISTIKA II., E-learningový prvek pro podporu výuky odborných a technických předmětů,v rámci projektu CZ.O4.01.3/3.2.15.2/0326, VŠB TU Ostrava, 2007, ISBN 978-80-248-1482-7.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **prof. Ing. Radim Briš, CSc.**

Datum zadání: 01.09.2013
Datum odevzdání: 07.05.2014

VYSOKÁ ŠKOLA BÁŇSKÁ-TECHNICKÁ UNIVERZITA Fakulta elektrotechniky a informatiky OSTRAVA

doc. RNDr. Jiří Bouchala, Ph.D.
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

V Ostravě 7. května 2014 .....

Ráda bych na tomto místě poděkovala prof. Ing. Radimu Brišovi, CSc., který mě v mé práci vedl a nasměroval mě vždy správným směrem, zaměstnancům Fakultní nemocnice v Ostravě, bez jejichž spolupráce by tato diplomová práce nemohla vzniknout, a své rodině, která mě během celého studia podporovala.

Abstrakt

Tato práce se věnuje Coxovu proporcionálnímu hazardnímu modelu a jeho aplikaci na data pacientů po operaci kolorekta, která poskytla Fakultní nemocnice v Ostravě (FNO). Cílem je sestavit vhodný Coxův proporcionální hazardní model a vytvořit program, který bude realizovat kroky k určení tohoto modelu. Při zpracování byly využity programy SPSS 20, R 3.0.2, RStudio 0.97.551.

Klíčová slova: analýza přežití, AFT modely, PH modely, Coxův proporcionální hazardní model, částečná věrohodnostní funkce, aproximace částečné věrohodnostní funkce, odhad základní hazardní funkce

Abstract

This work is about the Cox proportional hazard model and its application. The data of patients after colorectal surgery was provided by FNO. The goal is to study basics of survival analysis, fit the Cox proportional hazard model and create the program which will use the software R to construct the model. Other softwares that we use: SPSS, RStudio.

Keywords: survival analysis, AFT models, PH models, Cox proportional hazard model, partial likelihood function, approximation of partial likelihood function, estimate of baseline hazard function

Seznam použitých zkratek a symbolů

AR	–	arytmie
AFT	–	Accelerated Failure Time
FNO	–	Fakultní nemocnice Ostrava
GR	–	grading
MLE	–	Maximum Likelihood Estimation
PH model	–	Proportional Hazards model
ST	–	stadium
csv	–	Comma-separated values
html	–	HyperText Markup Language
$f(t)$	–	hustota pravděpodobnosti
$h(t)$	–	hazardní funkce
$\hat{h}(t)$	–	odhad hazardní funkce
$h_0(t)$	–	základní hazardní funkce
$\ell(\Theta\|x)$	–	věrohodnostní funkce
$D(t_{(j)})$	–	skupina jedinců, kteří zemřeli v čase $t_{(j)}$
DX	–	rozptyl náhodné veličiny X
EX	–	střední hodnota náhodné veličiny X
$F(t)$	–	distribuční funkce
$H(t)$	–	kumulativní hazardní funkce
$H_0(t)$	–	základní kumulativní hazardní funkce
$L(\Theta\|x)$	–	logaritmus věrohodnostní funkce
$P(X < t)$	–	pravděpodobnost, že náhodná veličina X je menší než t
$R(t_{(j)})$	–	riziková skupina v čase $t_{(j)}$
$S(t)$	–	funkce přežití
$\hat{S}(t)$	–	odhad funkce přežití
$S_0(t)$	–	základní funkce přežití

Obsah

Seznam tabulek

Seznam obrázků

1 Úvod

Cílem práce je se blíže seznámit s metodami používanými v analýze přežití. Popsat s jakým typem dat se v analýze přežití setkáváme, představit si nejpoužívanější modely a popsat modely typu AFT a PH.

Hlavním bodem pak je podrobněji rozebrat Coxův proporcionální model, kde nás konkrétně bude zajímat princip částečné věrohodnostní funkce, její odvození, aproximace, odhad jejich regresních koeficientů. Dále bychom se chtěli seznámit s metodami, které určují významnost těchto koeficientů v modelu. Také bychom se rádi podívali na to, co jsou to hazardní poměry a jak lze případně odhadnout základní hazardní funkci.

V neposlední řadě bychom chtěli aplikovat získané poznatky na datech pacientů po operaci kolorekta, které nám poskytla Fakultní nemocnice v Ostravě. Pro data týkající se doby do pooperačních komplikací bychom chtěli sestavit Coxův proporcionální hazardní model a na jeho základě ověřit domněnku, že by laparoskopická metoda měla být pro pacienta méně rizikovější, co se týká pooperačních komplikací, než metoda otevřená.

V rámci práce bychom chtěli analýzu poskytnutých dat podpořit programem, který by měl samostatně sestavit Coxův proporcionální model (tedy vyřešit problém s výběrem vhodných vysvětlujících proměnných do modelu) a na základě takto získaného modelu pak provést vyhodnocení dat v uživatelsky vhodném formátu.

2 Analýza přežití

Analýza přežití je rozsáhlé odvětví statistiky, které se zabývá analýzou doby trvání do výskytu jedné nebo více událostí. Využívá se téměř ve všech možných oborech a v rámci nich se upravuje i její terminologie. Proto např. v mechanice či ekonomii ji známe pod pojmem teorie spolehlivosti, v sociologii zase jako analýzu historie. Nejčastěji ji ovšem uplatňujeme v lékařském oboru, kde můžeme sledovat např. dobu do úmrtí u onkologických pacientů nebo dobu do výskytu pooperačních komplikací, což je zrovna náš případ. Díky ní můžeme např. určit, které sledované hodnoty u pacientů nejvíce ovlivňují vznik pooperačních komplikací a na základě těchto znalostí pak minimalizovat riziko jejich vzniku u nových pacientů.

2.1 S jakým typem dat pracujeme

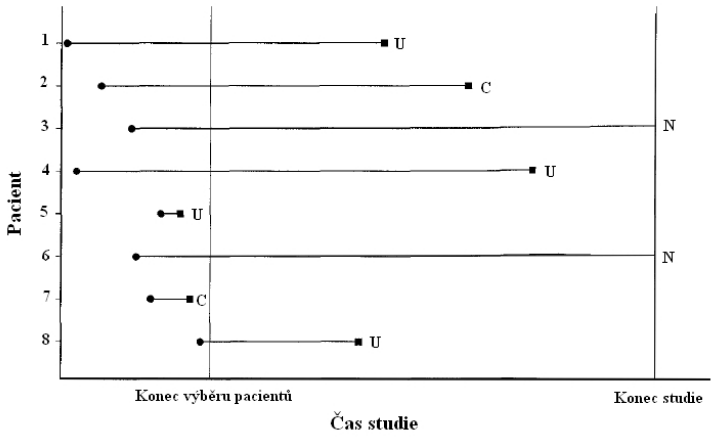
V praxi se často setkáváme s tzv. **neúplnými daty**, které také označujeme jako **cenzorovaná data**. Jak už název napovídá, jedná se o typ dat, u kterých přímo nepozorujeme dobu do pooperačních komplikací, ale pozorujeme údaj, který je neúplný/cenzorovaný. Cenzorování jsou následujících typů:

- **Cenzorování časem** (cenzorování 1. typu)
  Ke ztrátě dat dochází z toho důvodu, že doba do pooperačních komplikací u některých pacientů překročí dobu experimentu (studie). Doba experimentu T je stanovena předem a označujeme ji jako časový cenzor. Výsledkem je prvních r hodnot pořádkových statistik $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(r)} \leq T$ a fakt, že $X_{(r+1)} > T$, kde X značí náhodnou veličinu reprezentující dobu do pooperačních komplikací.

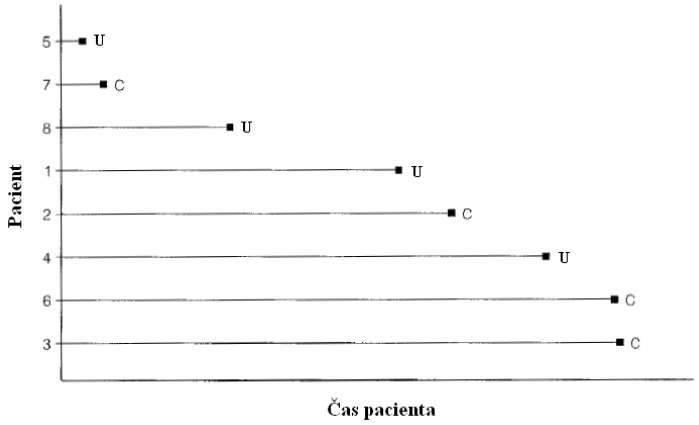
- **Cenzorování výskytem události** (cenzorování 2. typu)
  Zde je studie ukončena počtem pacientů u nichž dojde k pooperačním komplikacím (r). Tento počet stanovujeme na začátku a to tak, aby $r \leq n$, kde n je počet pacientů vstupujících do studie. Pozorování zahájíme v čase $t = 0$ a ukončíme v okamžiku, kdy dojde k pooperačním komplikacím r-tého pacienta. Výsledkem je pak prvních r hodnot pořádkových statistik $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(r)}$. Doba trvání studie (doba do pooperačních komplikací r-tého pacienta) je náhodná veličina $X_{(r)}$.

- **Náhodné cenzorování**
  Pokud se časové cenzory u jednotlivých pacientů liší (ukončování pozorování u jednotlivých pacientů je náhodné), mluvíme o **náhodném cenzorování**.
  Nechť X je náhodná veličina reprezentující dobu do pooperačních komplikací a T je náhodná veličina reprezentující časový cenzor. U každého pacienta pozorujeme buď X nebo T podle toho, zda dříve došlo k pooperačním komplikacím nebo zda dříve bylo sledování pacienta ukončeno. Výsledkem je pak n dvojic $(W_1, I_1), \ldots, (W_n, I_n)$, kde $W_i = min(X_i, T_i)$ a $I_i = 1$ pokud nedošlo k cenzorování ($W_i = X_i$), $I_i = 0$ pokud došlo k cenzorování ($W_i = T_i$).

Údaje bývají cenzorované z různých důvodů např. časových nebo ekonomických (finančně náročná studie), pacient nemusí být ochoten dále spolupracovat (odstěhuje se, může dojít k události z jiných než sledovaných příčin). [2]



Obrázek 1: Schéma času studie osmi pacientů



Obrázek 2: Schéma času pacientů

Obrázek 1 zachycuje schéma vstupu osmi pacientů do studie. Vidíme, že ne všichni pacienti vstupují do studie ve stejný čas. U pacientů 1,4,5 a 8 došlo k události (U), pacienti 2 a 7 byli ze studie vyřazeni - cenzorování (C), u pacientů 3 a 6 nedošlo k sledované události - konec studie (N). U pacientů 3 a 6 došlo tedy také k cenzorování, i když z jiných příčin než u pacientů 2 a 7.

Přestože pacienti vstupují do studie v různých časech (různá kalendářní období), vnímame jejich vstup stejně a to v čase t_0. Tuto skutečnost zachycuje obrázek 2.[3]

2.2 Základní funkce pro popis dat

Doba do pooperačních komplikací

Tímto termínem označujeme náhodnou veličinu T reprezentující dobu, která uplyne od začátku sledování pacienta do výskytu pooperačních komplikací.

Předpokládáme, že doba do pooperačních komplikací je nezáporná náhodná veličina se standardně definovanou **distribuční funkcí** $F(t)$. Tato funkce je neklesající funkcí času a vyjadřuje pravděpodobnost, že v intervalu $(0, t)$ dojde k pooperačním komplikacím. V případě, že je distribuční funkce spojitá, můžeme obdobně zadefinovat i **hustotu pravděpodobnosti** $f(t)$.

Funkce přežití

S distribuční funkcí úzce souvisí pojem **funkce přežití**, která vyjadřuje pravděpodobnost, že v intervalu $(0, t)$ nedojde ke sledované události (např. úmrtí, pooperační komplikace). Definujeme ji jako

$$S(t) = P(T \geq t) = 1 - F(t).$$

Na rozdíl od distribuční funkce je $S(t)$ nerostoucí funkcí času.

Hazardní funkce

Pro dobu do pooperačních komplikací T se spojitým rozdělením popsaným distribuční funkcí $F(t)$ a hustotou pravděpodobnosti $f(t)$ definujeme **hazardní funkci**

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}, \quad \text{pro } F(t) \neq 1 \tag{2.1}$$

Mluvíme-li o hazardní funkci, musíme si uvědomit, že se nejedná o pravděpodobnost, ale o poměr hustoty pravděpodobnosti a funkce přežití. Pravděpodobnost, že dojde k pooperační komplikaci v následujícím krátkém úseku délky Δt za předpokladu, že k ní do času t nedošlo, je přibližně $h(t) \cdot \Delta t$.

Kumulativní hazardní funkce

Můžeme se setkat i s pojmem **kumulativní hazardní funkce**, která je definovaná

$$H(t) = \int_0^t h(u)du = -ln(S(t)) + ln(S(0)) = -ln(S(t)) \tag{2.2}$$

Mezi $f(t), F(t), S(t), h(t)$ existují vzájemné převodní vztahy (Tabulka 1). Například z výrazu (2.1) snadno odvodíme

$$h(t) = \frac{f(t)}{S(t)} = \frac{dF(t)}{dt} \cdot \frac{1}{S(t)} = \frac{d(1 - S(t))}{dt} \cdot \frac{1}{S(t)} = -\frac{dS(t)}{dt} \cdot \frac{1}{S(t)} = -\frac{dlnS(t)}{dt} \tag{2.3}$$

S využitím zřejmé vlastnosti, že $S(0) = 1$ [**pozn**. plyne ze vztahu $S(t) = 1 - F(t)$, $F(0) = 0$] dostaneme vztah

$$S(t) = e^{-\int_0^t h(u)du} \tag{2.4}$$

A hustota pravděpodobnosti

$$f(t) = h(t) \cdot e^{-\int_0^t h(u)du} \tag{2.5}$$

[12]

	$S(t)$	$F(t)$	$f(t)$	$h(t)$
$S(t)$		$1 - F(t)$	$1 - \int_0^t f(u)du$	$e^{\left[-\int_0^t h(u)du\right]}$
$F(t)$	$1 - S(t)$		$\int_0^t f(u)du$	$1 - e^{\left[-\int_0^t h(u)du\right]}$
$f(t)$	$-\frac{dS(t)}{dt}$	$\frac{dF(t)}{dt}$		$h(t) \cdot e^{\left[-\int_0^t h(u)du\right]}$
$h(t)$	$-\frac{dlnS(t)}{dt}$	$-\frac{dln(1-F(t))}{dt}$	$f(t)\left(1 - \int_0^t f(u)du\right)^{-1}$	

Tabulka 1: Matematické převodní vztahy [2]

Věrohodnostní funkce

Věrohodnostní funkce je pro nás v tuto chvíli jedna z nejdůležitějších funkcí vůbec, a to hlavně proto, že se kolem ní točí celý základ teorie Coxova proporcionálního hazardního modelu, s kterým se v rámci této práce chceme dopodrobna seznámit.

Definice 2.1 *Nechť $X = (X_1, \ldots, X_n)$ je náhodný výběr a $x = (x_1, \ldots, x_n)$ je jeho realizace. Nechť je dále popsána populace pomocí regulární hustoty $f(x, \Theta)$, kde Θ je neznámý parametr. Potom funkci*

$$\ell(\Theta|x) = \ell(\Theta|x_1, \ldots, x_n) = f(x_1, \Theta)f(x_2, \Theta)\ldots f(x_n, \Theta) = \prod_{i=1}^{n} f(x_i, \Theta) \tag{2.6}$$

*budeme nazývat **věrohodnostní funkcí**. [13]*

Pojem regulární hustota je zadefinován v [2]. Pokud mluvíme o rozdělení, považujeme parametr za fixní a pozorování se mění. Pokud mluvíme o věrohodnosti, pak jsou pozorování fixní a parametr se může měnit. Věrohodnostní funkci můžeme interpretovat jako

pravděpodobnost toho, že při daném parametru Θ opět naměříme stejné hodnoty. $\ell(\Theta|x)$ je základem MLE (metoda maximální věrohodnosti), která se využívá v PH modelu. Zdůrazním, že pokud $X_1, X_2, \ldots, X_n$ je množina nezávislých náhodných veličin z rozdělení s hustotami $f_i(x_i, \Theta)$, $i = 1, \ldots, n$, pak věrohodnostní funkce může být získána jako součin těchto hustot.

Zpravidla nepoužíváme přímo věrohodnostní funkci, ale pracujeme s jejím logaritmem (2.7), protože jsou s ním výpočty jednodušší, jak si ukážeme později.

$$L(\Theta|x) = \sum_{i=1}^{n} ln(f(x_i, \Theta)) \tag{2.7}$$

2.3 Další charakteristiky a používané pojmy

Střední doba do pooperačních komplikací

Střední doba do pooperačních komplikací se definuje jako střední hodnota náhodné veličiny, tj. doby do pooperačních komplikací T

$$ET = \int_0^\infty t f(t) dt \tag{2.8}$$

Střední dobu do pooperačních komplikací lze spočítat z funkce přežití $S(t)$, a to díky *vlastnosti*: Nechť nezáporná náhodná veličina T má funkci přežití $S(t)$. Nechť $ET^k < +\infty$, kde k je přirozené číslo (existují konečné obecné momenty všech řádů). Pak

$$ET^k = k \int_0^\infty t^{k-1} S(t) dt \tag{2.9}$$

Pro střední dobu do pooperačních komplikací ($k = 1$) pak dostáváme

$$ET = \int_0^\infty S(t) dt \tag{2.10}$$

Pro rozptyl doby do pooperačních komplikací platí

$$DT \quad = \quad ET^2 - (ET)^2 = 2 \int_0^\infty tS(t) dt - \left(\int_0^\infty S(t) dt \right)^2 \tag{2.11}$$

Riziková skupina

Protože se často budeme setkávat s pojmem **rizikové skupiny**, podíváme se, co se pod tímto názvem ukrývá.

Nechť máme následující pozorované časy do pooperačních komplikací $t_1 < t_2 < \cdots < t_m$. Do rizikové skupiny v čase t_j, $j = 1, \ldots, m$, kterou označujeme jako $R(t_j)$, patří ti jedinci, jejichž časy jsou rovny nebo větší než t_j. Počet jedinců v této skupině značíme n_j.

Příklad 2.1
Mějme u osmi pacientů zaznamenané tyto časy do sledované události (sledovanou událostí může být např. úmrtí, pooperační komplikace aj.): $3, 5, 5+, 7, 9+, 10, 11, 15+$. Kde $+$ značí cenzorovaný údaj, který chápeme tak, že doba do pooperačních komplikací u daného pacienta je větší než tento zaznamenaný čas.

Do rizikové skupiny v čase $t_j = 5$ budou patřit pacienti s následujícími časy do výskytu události

$$\begin{aligned} R(5) &= \{5, 5+, 7, 9+, 10, 11, 15+\} \\ n_5 &= 7 \end{aligned}$$

Namátkou další skupiny budou např.

$$\begin{aligned} R(7) &= \{7, 9+, 10, 11, 15+\} \\ n_7 &= 5 \\ R(10) &= \{10, 11, 15+\} \\ n_{10} &= 3 \end{aligned}$$

■

3 Modely v analýze přežití

Úplně v nejjednodušším případě můžeme vytvořit studii, v rámci které budeme sledovat pouze doby do pooperačních komplikací. Charakter této náhodné veličiny (dobu do pooperačních komplikací) se pak budeme snažit zachytit pomocí nějakého modelu. Na výběr máme různé přístupy k tomuto problému. Pokud jsme například podobnou studii už realizovali v minulosti, můžeme již mít představu o tom, jak se tato náhodná veličina chová. Takovou informaci následně můžeme využít při sestavování modelu. Hovoříme o tzv. **parametrických** modelech, které jsou založené na předpokladu, že už něco víme o rozdělení dané náhodné veličiny. Většinou takové štěstí ale nemáme a musíme se spolehnout na tzv. **neparametrické** modely, které vycházejí pouze z naměřených hodnot a nemusejí být tak přesné. A aby to nebylo příliš jednoduché, existují ještě tzv. **semiparametrické** modely, které jsou kombinací dvou předchozích přístupů, a které nás budou nejvíce zajímat, protože do této skupiny řadíme již zmíněný Coxův model, který budeme chtít podrobněji rozebrat.

Intuitivně sledovat pouze dobu do pooperačních komplikací nebude úplně dostačující. Svůj vliv budou mít určitě i jiné faktory. Např. pacient v pokročilém stádiu nemoci bude mít nejspíše menší šanci na hladký průběh léčby než pacient, který je na tom zdravotně lépe. **Regresní model** nám umožní právě do modelu takovéto údaje zanést. Existují dva možné přístupy, jak model vytvořit. Hovoříme tak pak buď o **AFT modelu** (*Accelerated Failure Time*) nebo o **PH modelu** (*Proportional Hazard*). Zatímco první reflektuje vliv vstupních proměnných jako modifikaci času, druhý tento vliv reflektuje jako modifikaci rizika. Co to znamená si nyní ukážeme.

3.1 AFT modely

Nechť u i-tého pacienta máme naměřené hodnoty $(t_i, c_i, x_1, \ldots, x_p)$, $i = 1, \ldots, n$, kde t_i je zaznamenaná doba do pooperačních komplikací, c_i dává informaci o tom, zda-li se jedná o cenzorovaný údaj či nikoliv, $x_1, \ldots, x_p$ jsou vysvětlující proměnné (např. váha, výška, tlak aj.), které by mohly mít vliv na dobu do pooperačních komplikací. AFT modely definují model pro dobu do pooperačních komplikací T následovně

$$\begin{aligned} Y = ln(T) &= \beta_0 + \beta' x + \sigma\epsilon, \\ T &= e^{\beta_0 + \beta' x + \sigma\epsilon} = e^{\beta_0 + \sigma\epsilon} \cdot e^{\beta' x} \end{aligned} \tag{3.1}$$

kde β_0 je absolutní člen, $\beta = (\beta_1, \ldots, \beta_p)$ je vektor neznámých koeficientů, $x = (x_1, \ldots, x_p)$ je vektor vysvětlujících proměnných a ϵ je chybová složka, která má příslušné log-rozdělení (např. jedná-li se o exponencionální rozdělení má log-exponencionální rozdělení a $\sigma = 1$).

U pacientů s vektorem vysvětlujících proměnných $x = 0$ definujeme tzv. **referenční dobu** do pooperačních komplikací

$$T_0 = e^{\beta_0 + \sigma\epsilon}.$$

Dosadíme-li tento vztah do (3.1) můžeme vyjádřit

$$T = T_0 \cdot e^{\beta' x}. \tag{3.2}$$

Vidíme tedy, že vliv vysvětlujících proměnných je v modelu zahrnut jako modifikace referenční doby do pooperačních komplikací.

Ještě výstižněji můžeme vysvětlit podstatu AFT modelů, pokud se podíváme na funkci přežití.

$$S(t,x) = P(T > t|x) = P(T_0 e^{\beta' x} > t) = P(T_0 > te^{-\beta' x}) = S_0(te^{-\beta' x}), \tag{3.3}$$

kde $S_0(t) = P(T_0 > t)$ se označuje jako **základní funkce přežití**. Vidíme, že pravděpodobnost toho, že pacient s naměřenými hodnotami x nebude mít v intervalu $(0,t)$ pooperační komplikace, je stejná jako pravděpodobnost, že pacient z referenční skupiny (pacient s naměřenými hodnotami $x = 0$) nebude mít pooperační komplikace v intervalu $(0, te^{-\beta' x})$.

Konkrétně pokud budeme mít případ s pouze jednou vysvětlující proměnnou, která může nabývat hodnoty buď $x = 0$ nebo $x = 1$ a bude-li hodnota koeficientu $\beta = ln(2)$, pak $S(t, x = 1) = S_0(t \cdot 2)$. Což můžeme slovně interpretovat, že pravděpodobnost toho, že pacient s naměřenou hodnotou $x = 1$ nebude mít do času t pooperační komplikace je stejná, jako pravděpodobnost, že pacient z referenční skupiny nebude mít pooperační komplikace po dobu dvakrát delší. [24]

3.2 PH modely

Další skupinou modelů jsou **proporcionálně hazardní modely**. PH modely respektují hazardní funkci, jejichž podobu dále ovlivňují vysvětlující proměnné. Např. užívání léku může snížit míru rizika cévní mozkové příhody na polovinu, nebo změna materiálu, ze kterého je vyrobena součástka, může zdvojnásobit riziko selhání.

Přítomnost cenzorovaných časů v datech dělá analýzu přežití zajímavější, bohužel nelze na analýzu těchto dat využít klasické statistické metody a je třeba sáhnout po sofistikovanějších modelech, které dokážou tyto neúplná data zohlednit. Jeden z takových modelů je **Coxův model proporcionálního rizika** (dále zkráceně **Coxův model** či se pro něj užívá přímo výrazu **PH model**). Abychom zcela pochopili princip PH modelů, začneme od nejtriviálnějších faktů.

Funkce, která charakterizuje čas přežití a nejlépe dokáže zachytit proces stárnutí, je hazardní funkce $h(t)$. V nejjednodušším případě může být $h(t)$ konstantní

$$h(t) = \Theta_0. \tag{3.4}$$

Odhad parametru lze provést snadněji, pokud umožníme, aby nabýval libovolné hodnoty. Avšak v případě (3.4) musí být parametr větší nebo roven nule, protože hazardní funkce je striktně pozitivní.

Jedním možným způsobem, jak zachovat obě požadované vlastnosti ($h(t) > 0$ a $\Theta_0 \in \langle -\infty, +\infty \rangle$), je parametrizovat hazardní funkci následovně

$$h(t) = e^{\beta_0}, \tag{3.5}$$

kde $\beta_0 = ln(\Theta_0)$ a je tedy neomezené. Přirozeným způsobem, jak zahrnout i proměnné, je jejich ponechání v logaritmickém měřítku. Speciálně pro proměnnou x je log-hazardní funkce

$$ln\left[h(t)\right] = \beta_0 + \beta_1 x, \tag{3.6}$$

resp. pokud máme p vysvětlujících proměnných $x = (x_1, \ldots, x_p)$ a p příslušných neznámých koeficientů $\beta = (\beta_1, \cdots, \beta_p)$

$$ln\left[h(t)\right] = \beta_0 + \beta' x$$

a hazardní funkce

$$h(t) = e^{\beta_0 + \beta_1 x}, \quad \text{resp.} \quad h(t) = e^{\beta_0 + \beta' x}. \tag{3.7}$$

Skutečnost, že hazardní funkce (3.7) neobsahuje časovou proměnnou, nemusí být v některých situacích ideální. Např. lidský život se řídí Weibullovým rozdělením (”vanová křivka”). Je ale možné charakterizovat hazard jako explicitní funkcí času i vstupních proměnných. V podstatě plně parametrizovaná hazardní funkce dosahuje dvou cílu současně:

(1) Popisuje základní distribuci času přežití (chybová část).

(2) Charakterizuje, jak se změní distribuce jako funkce proměnných (systematická část).

Jeden možný tvar regresního modelu hazardní funkce je tedy

$$h(t, x, \beta) = h_0(t) r(x, \beta). \tag{3.8}$$

V případě (3.8) je hazardní funkce výsledkem součinu dvou funkcí:

- $h_0(t)$ - popisuje, jak se hazardní funkce mění v čase
- $r(x, \beta)$ - popisuje, jak se hazardní funkce mění v závislosti na vstupních proměnných

Tyto funkce musí být zvoleny tak, aby bylo splněno $h(t, x, \beta) > 0$. Pokud $r(x, \beta) = 1$, pak $h_0(t)$ je hazardní funkcí. V případě, že $r(x, \beta)$ je parametrizována tak, že $r(x = 0, \beta) = 1$, nazýváme $h_0(t)$ **základní hazardní funkcí**. Na základě modelu (3.8) je poměr hazardních funkcí HR (z angl. *Hazard Ratio*) dvou jedinců se vstupními proměnnými označenými jako x_1 a x_0 dán vztahem

$$\begin{aligned} HR(t, x_1, x_0) &= \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)} \\ HR(t, x_1, x_0) &= \frac{h_0(t) r(x_1, \beta)}{h_0(t) r(x_0, \beta)} = \frac{r(x_1, \beta)}{r(x_0, \beta)} \end{aligned} \tag{3.9}$$

Vidíme, že **hazardní poměr** závisí pouze na funkci $r(x, \beta)$. Pokud $HR(t, x_1, x_0)$ má jasnou klinickou interpretaci, není tvar základní hazardní funkce $h_0(t)$ tak důležitý. Toto je ta podstatná vlastnost **semiparametrických modelů**. Není třeba specifikovat $h_0(t)$. Zároveň

si všimněme, že hazardní poměr se v čase nemění, je konstantní. Tento fakt je znám jako **podmínka proporcionality rizik**. Jednoduše riziko pooperačních komplikací jednoho pacienta je konstantním násobkem rizika pooperačních komplikací druhého pacienta.

Cox (1972) byl první, který navrhnul model (3.9), kde použil funkci $r(x, \beta) = e^{\beta' x}$. Takto parametrizovaná hazardní funkce, pak má tvar

$$h(t, x, \beta) = h_0(t) e^{\beta' x} \tag{3.10}$$

a hazardní poměr je

$$HR(t, x_1, x_0) = e^{\beta'(x_1 - x_0)} \tag{3.11}$$

Další otázkou je, jak vypadá funkce přežití jednak obecně pro PH model a jednak pro Coxův model (3.10). Na základě vztahů z tabulky (1) a kumulativní hazardní funkce můžeme psát

$$S(t, x, \beta) = e^{-H(t,x,\beta)} \tag{3.12}$$

V případě, že je doba přežití absolutně spojitá, můžeme psát

$$\begin{aligned} H(t, x, \beta) &= \int_0^t h(u, x, \beta) du \\ &= r(x, \beta) \int_0^t h_0(u) du \\ &= r(x, \beta) H_0(t). \end{aligned} \tag{3.13}$$

Po dosazení vztahu (3.13) do vztahu (3.12) dostáváme

$$\begin{aligned} S(t, x, \beta) &= e^{-r(x,\beta) H_0(t)} \\ S(t, x, \beta) &= \left[e^{-H_0(t)} \right]^{r(x,\beta)} \\ &= [S_0(t)]^{r(x,\beta)}, \end{aligned} \tag{3.14}$$

kde $S_0(t) = e^{-H_0(t)}$ nazýváme **základní funkci přežití**.

Pro Coxův model má pak funkce přežití tvar

$$S(t, x, \beta) = [S_0(t)]^{e^{\beta' x}} \tag{3.15}$$

[10]

3.3 Parametrické modely

Jak už jsme naznačili, parametrické modely vychází z předpokladu, že víme, jakému rozdělení podléhají zpracovávaná data. Můžeme tak vytvořit např. exponencionální regresní model, Weibullův regresní model aj. Některé modely lze zapsat jak ve tvaru odpovídajícímu AFT modelu, tak ve tvaru odpovídajícímu PH modelu. A naopak u některých modelů lze získat pouze jeden ze zmíněných zápisů. Co tím myslíme, si ukážeme nyní.

Exponencionální regresní model

Předpokládejme, že náhodná veličina T (doba do pooperačních komplikací) se řídí exponencionálním rozdělením $E(\lambda)$. Víme, že funkce přežití a odpovídající hazardní funkce exponencionálního rozdělení $E(\lambda)$ jsou dány vztahy

$$S(t) = e^{-\lambda t}, \qquad h(t) = \lambda. \tag{3.16}$$

Neznámý parametr λ můžeme vyjádřit jako funkci p vysvětlujících proměnných $\lambda = e^{-(\beta_0+\beta_1 x+\cdots+\beta_p x_p)}$. V případě jedné vysvětlující proměnné bude $\lambda = e^{-(\beta_0+\beta_1 x)}$. Pro jednoduchost bez újmy na obecnosti předpokládejme, že pracujeme pouze s jednou vysvětlující proměnnou. Funkci přežití $S(t)$ a základní funkci přežití $S_0(t)$ lze přepsat do tvaru

$$S(t) = e^{-te^{-(\beta_0+\beta_1 x)}}, \qquad S_0(t) = e^{-te^{-\beta_0}}. \tag{3.17}$$

Příslušná hazardní a základní hazardní funkce

$$h(t) = e^{-(\beta_0+\beta_1 x)}, \qquad h_0(t) = e^{-\beta_0}, \tag{3.18}$$

což odpovídá skutečnosti, že hazardní funkce exponencionálního rozdělení je v čase konstantní.

Tvar funkce přežití $S(t)$, který je dán vztahem (3.16), můžeme přepsat do podoby odpovídající AFT modelům nebo do podoby odpovídající PH modelům.

AFT tvar funkce přežití
S využitím vztahů (3.17) dostáváme

$$S(t) = e^{-te^{-(\beta_0+\beta_1 x)}} = e^{-(te^{-\beta_1 x})\cdot(e^{-\beta_0})} = S_0(te^{-\beta_1 x}),$$

což odpovídá předpisu AFT modelu.

PH tvar funkce přežití
Obecně jsme odvodili, že pro funkci přežití platí vztah (3.14). Uvědomíme-li si, že na základě (3.18) můžeme hazardní funkci vyjádřit jako

$$h(t) = e^{-(\beta_0+\beta_1 x)} = e^{-\beta_0}e^{-\beta_1 x} = h_0(t)e^{-\beta_1 x},$$

je jasné, že naší funkcí $r(x, \beta)$ je právě $e^{-\beta_1 x}$. Což plně odpovídá odvození PH tvaru funkce přežití

$$S(t) = e^{-te^{-(\beta_0+\beta_1 x)}} = e^{-te^{-\beta_0}e^{-\beta_1 x}} = \left(e^{-te^{-\beta_0}}\right)^{e^{-\beta_1 x}} = (S_0(t))^{e^{-\beta_1 x}}$$

Weibullův regresní model

Předpokládejme, že náhodná veličina T (doba do pooperačních komplikací) se řídí Weibullovým rozdělením $W(\Theta, \alpha)$. Funkce přežití a odpovídající hazardní funkce Weibullova rozdělení $W(\Theta, \alpha)$, kde $\Theta = 1/\lambda$ se nazývá parametr měřítka a α parametr tvaru, jsou

$$S(t) = e^{-(\lambda t)^\alpha}, \qquad h(t) = \alpha\lambda^\alpha t^{\alpha-1} \tag{3.19}$$

Parametr λ můžeme opět vyjádřit jako funkci vysvětlujících proměnných a v případě jedné vysvětlující proměnné x tak dostáváme

$$S(t) = e^{-(te^{-(\beta_0+\beta_1 x)})^\alpha}, \qquad S_0(t) = e^{-(te^{-\beta_0})^\alpha} = e^{-t^\alpha e^{-\beta_0 \alpha}} \tag{3.20}$$

a hazardní funkce

$$h(t) = \frac{\alpha t^{\alpha-1}}{(e^{\beta_0+\beta_1 x})^\alpha} \qquad h_0(t) = \frac{\alpha t^{\alpha-1}}{e^{\beta_0 \alpha}} \tag{3.21}$$

Zatímco v exponencionálním regresním modelu platí, že $\sigma = 1$, vztah mezi parametrem α a parametrem σ ve Weibullovém regresním modelu je $\alpha = 1/\sigma$.

AFT tvar funkce přežití

$$S(t) = e^{-(te^{-(\beta_0+\beta_1 x)})^\alpha} = e^{-t^\alpha e^{-\beta_0 \alpha} e^{-\beta_1 x \alpha}} = S_0(te^{-\beta_1 x}),$$

což odpovídá předpisu AFT modelu.

PH tvar funkce přežití
Uvědomíme-li si, že můžeme hazardní funkci vyjádřit jako

$$h(t) = \frac{\alpha t^{\alpha-1}}{(e^{\beta_0+\beta_1 x})^\alpha} = h_0(t) e^{-\beta_1 x \alpha}$$

je jasné, že naší funkcí $r(x, \beta)$ je právě $e^{-\beta_1 x \alpha}$. Což plně odpovídá odvození PH tvaru funkce přežití

$$S(t) = e^{-(te^{-(\beta_0+\beta_1 x)})^\alpha} = e^{-t^\alpha e^{-\beta_0 \alpha} e^{-\beta_1 x \alpha}} = (S_0(t))^{e^{-\beta_1 x \alpha}}$$

Log-normální regresní model

Předpokládejme, že náhodná veličina T (doba do pooperačních komplikací) se řídí log-normálním rozdělením $LN(\mu, \sigma)$. Chybová složka regresního modelu se bude řídit normálním rozdělením $N(0, \sigma)$. Protože hazardní funkce log-normálního rozdělení má velmi specifický tvar, není splněn předpoklad proporcionality rizik a hazardní poměr proto nebude konstantní pro všechny doby do pooperačních komplikací. Jsme tedy schopni vytvořit pouze AFT model tohoto rozdělení. Víme, že funkce přežití je dána vztahem

$$S(t) = 1 - \Phi\left(\frac{ln(t) - \mu}{\sigma}\right), \tag{3.22}$$

kde Φ je distribuční funkce normovaného normálního rozdělení. Pokud máme pro jednoduchost jednu vysvětlující proměnnou a vyjádříme parametr μ jako $\mu = \beta_0 + \beta_1 x$, dostaneme

$$S(t) = 1 - \Phi\left(\frac{ln(t) - (\beta_0 + \beta_1 x)}{\sigma}\right), \qquad S_0(t) = 1 - \Phi\left(\frac{ln(t) - \beta_0}{\sigma}\right) \tag{3.23}$$

AFT tvar funkce přežití

$$S(t) = 1 - \Phi\left(\frac{ln(t) - (\beta_0 + \beta_1 x)}{\sigma}\right) = 1 - \Phi\left(\frac{ln(te^{-\beta_1 x}) - \beta_0}{\sigma}\right) = S_0(te^{-\beta_1 x}),$$

což odpovídá předpisu AFT modelu.

Další model, pro který dokážeme vytvořit pouze AFT, je např. **zobecněný gamma regresní model**. Více k tomuto modelu lze nalézt v [24], [25].

3.4 Neparametrické modely

Neparametrické modely většinou použijeme tehdy, pokud nevíme nic o rozdělení ze kterého pocházejí data. Na ukázku si představíme dvě metody: první nám umožní odhadnout funkci přežití, druhá pak hazardní funkci. Tyto dvě metody v sobě nezahrnují vliv vysvětlujících proměnných. Odhad se provádí na základě pozorovaných dob do pooperačních komplikací.

3.4.1 Kaplan-Meierův odhad funkce přežití

Kaplan-Meierův odhad nám umožňuje na základě pozorovaných dat odhadnout funkci přežití. Není mým záměrem odvozovat dopodrobna konstrukci tohoto odhadu, a proto přejdu rovnou na praktický příklad, který objasní, jak můžeme z cenzorovaných dat získat odhad funkce přežití $\hat{S}(t)$.

Příklad 3.1
Proveďte Kaplan-Meierův odhad, pokud pozorované časy do pooperačních komplikací byly $1, 2, 2, 2+, 3, 3+, 4, 4+, 5+, 6, 7, 8, 9, 9, 10$.

Prvně si všimneme, že máme data již uspořádaná. Celkový počet jedinců ve studii $n = 15$. Dále musíme určit pozorované odlišné necenzorované časy $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$, které jsou v našem případě $\{1, 2, 3, 4, 6, 7, 8, 9, 10\}$, $m = 9$. Pro výpočet použijeme vzorec

$$\hat{S}(t_{(k)}) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right),$$

kde n_j je počet jedinců v rizikové skupině $R(t_{(j)})$ a d_j je počet jedinců ve skupině $D(t_{(j)})$, která označuje jedince, kteří mají časy pooperačních komplikací rovny času $t_{(j)}$. Spočí-

tejme prvních pár hodnot:

$$\begin{aligned}
\hat{S}(1) &= \left(\frac{15-1}{15}\right) = 0.933 \\
\hat{S}(2) &= \hat{S}(1) \cdot \left(\frac{14-2}{14}\right) = 0.8 \\
\hat{S}(3) &= \hat{S}(2) \cdot \left(\frac{11-1}{11}\right) = 0.727 \\
\hat{S}(4) &= \hat{S}(3) \cdot \left(\frac{9-1}{9}\right) = 0.646 \\
\hat{S}(6) &= \hat{S}(4) \cdot \left(\frac{6-1}{6}\right) = 0.539
\end{aligned}$$

V tabulce 2 můžeme vidět, jaké výsledky dává statistický software **R**. Příslušný odhad funkce přežití $\hat{S}(t)$ je pak graficky znázorněn na obrázku 3.

čas t	**počet v ohrožení**	**počet událostí**	**Ŝ(t)**
1	15	1	0.933
2	14	2	0.800
3	11	1	0.727
4	9	1	0.646
6	6	1	0.539
7	5	1	0.431
8	4	1	0.323
9	3	2	0.108
10	1	1	0.000

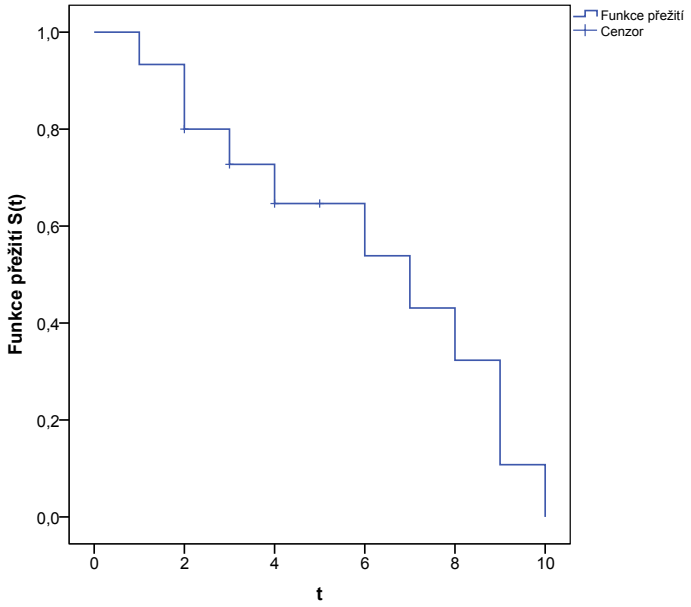
Tabulka 2: Kaplan-Meierův odhad (výstupní tabulka v softwaru **R**)

■

Kromě Kaplan-Meierova odhadu funkce přežití, který patří mezi nejznámější, lze pro odhad funkce přežití použít metodu **Life-table** [10] nebo **Nelson-Aalenův** odhad, který je velmi dobře popsán v [3] v kapitole *2.1.3*.

3.4.2 Kaplan-Meierův odhad hazardní funkce

Intuitivně můžeme ze získaných dat odhadnout hazardní funkci jako podíl počtu pooperačních komplikací v daném čase ku počtu jedinců v rizikové skupině v daném čase. Pokud předpokládáme, že je hazardní funkce mezi jednotlivými pozorovanými časy pooperačních komplikací konstantní, můžeme získat hazard za jednotku času, jestliže daný výraz navíc podělíme délkou časového intervalu. Takže pokud si označíme d_j počet pooperačních komplikací v čase $t_{(j)}$, $j = 1, 2, \ldots, m$ a n_j počet jedinců v rizikové skupině



Obrázek 3: Kaplan-Meierův odhad funkce přežití př. 3.1

$R(t_{(j)})$, můžeme hazardní funkci v intervalu $t_{(j)}$ a $t_{(j+1)}$ odhadnout jako

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, \tag{3.24}$$

pro $t_{(j)} \leq t < t_{(j+1)}$ a $\tau_j = t_{(j+1)} - t_{(j)}$. Je jasné, že není možné v intervalu, který začíná v posledním největším pozorovaném čase pooperačních komplikací $t_{(m)}$, použít pro odhad hazardu vztah (3.24), protože tento interval je otevřený.

Odhad (3.24) se nazývá *Kaplan-Meierův*, protože funkce přežití z něj odvozená odpovídá právě Kaplan-Meierovu odhadu. Lehce to zdůvodníme, pokud si uvědomíme, že $\hat{h}(t)$, $t_{(j)} \leq t < t_{(j+1)}$ je odhad rizika pooperačních komplikací za jednotku času v j-tém intervalu. Pravděpodobnost pooperačních komplikací v tomto intervalu je tedy $\hat{h}(t)\tau_j$, což je d_j/n_j. Proto odhad odpovídající pravděpodobnosti přežití v daném intervalu bude $1 - (d_j/n_j)$ a odhad funkce přežití bude stejný jako v kapitole (3.4.1).

3.5 Semiparametrické modely

Jak už bylo zmíněno dříve, semiparametrické modely jsou kombinací parametrického a neparametrického přístupu. Jejich největší výhodou je, že nemusíme znát tvar základní hazardní funkce. Mezi nejznámější semiparametrický model patří Coxův model proporcionálních rizik, který má tvar

$$h(t) = h_0(t)e^{\beta' x},$$

kde β je vektor koeficientů $\beta = (\beta_1, \ldots, \beta_p)$ a x je vektor vysvětlujících proměnných $x = (x_1, \ldots, x_p)$.

3.6 Základní metody používané v modelování

Než přistoupíme ke kapitole, která se bude podrobněji věnovat Coxovu proporcionálnímu modelu, seznámíme se pro přehled s některými metodami, které využíváme v analýze přežití.

3.6.1 Metoda maximální věrohodnosti

Přirozeným způsobem, jak odhadnout neznámý parametr Θ, je maximalizovat věrohodnostní funkci. Avšak použít k této maximalizaci přímo věrohodnostní funkci se ukazuje nepraktické. Je jednodušší maximalizovat logaritmus této funkce. Obě funkce, věrohodnostní $\ell(\Theta|x)$ i její logaritmus $L(\Theta|x)$, dosahují maxima v témže bodě, takže tato úprava má smysl.

Ukažme si podstatu MLE na jednoduchém diskrétním případě.

Příklad 3.2
Nechť máme výběr (X_1, X_2, X_3, X_4) z alternativního rozdělení s neznámým parametrem p (tj. $P(X = 1) = p$, $P(X = 0) = 1 - p$). Založme odhad parametru p na naměřených datech, a ty jsou $(0, 0, 1, 0)$. Pro jednoduchost předpokládejme, že parametr p má hodnotu buď $p = 0.2$ nebo $p = 0.8$. Pro pravděpodobnost pozorovaných dat z alternativního rozdělení platí:

$$P(X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 0) = p(1 - p)^3,$$

což se pro $p = 0.2$ rovná hodnotě 0.1024, pro $p = 0.8$ pak 0.0064. Metoda MLE spočívá v tom, že za odhad parametru p vezmeme tu hodnotu, pro kterou je pravděpodobnost takto naměřených dat největší, tedy $p = 0.2$. ■

Na dalším příkladě si ukážeme typické použití MLE.

Příklad 3.3
Odhadněte metodou maximální věrohodnosti parametr μ normálního rozdělení $N(\mu, \sigma^2)$ s distribuční funkcí

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right),$$

kde parametr σ^2 známe. Neznámým parametrem Θ je v tomto případě μ. Pro metodu maximální věrohodnosti tak dostáváme

$$\begin{aligned} L(\mu|X_1, X_2, \ldots, X_n) &= ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right), \text{ tedy} \\ L(\mu|X_1, X_2, \ldots, X_n) &= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \end{aligned} \tag{3.25}$$

Dále budeme hledat maximum funkce (3.25) - zderivujeme ji podle neznámého parametru μ a položíme rovno 0.

$$\begin{aligned}\frac{\partial L(\Theta|X_1, X_2, \ldots, X_n)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) &= 0 \\ \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\mu} &= \bar{\mathrm{X}}\end{aligned}$$

Zjistili jsme, že MLE odhadem střední hodnoty normálního rozdělení je výběrový průměr. ■

V případě, že nebudeme znát ani parametr σ^2, budeme logaritmus věrohodnostní funkce $L(\Theta|x)$ derivovat jednak podle proměnné μ, tak i podle proměnné σ^2. Obě rovnice položíme rovny 0 a vyřešíme soustavu dvou rovnic o dvou neznámých. [2]

3.6.2 Srovnání přežití dvou skupin pomocí Log-rank testu

Nejjednodušším způsobem, jak srovnat přežití dvou skupin, je vykreslit si odpovídající funkce přežití těchto skupin do jednoho grafu. Odtud pak můžeme získat potřebné informace. Bohužel nám ale takový postup nedá odpověď na otázku, jestli jsou rozdíly mezi dvěma skupinami významné či nikoliv. Proto zde představím *Log-rank test*, který tento problém řeší velmi elegantním způsobem.

Existuje mnoho variací tohoto testu, avšak všechny jsou založené na kontingenční tabulce 3, která se konstruuje pro každý daný čas přežití $t_{(j)}$.

Událost/Skupina	**1**	**2**	**Celkem**
Počet úmrtí	d_{1j}	d_{2j}	d_j
Počet přežití	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Počet v rizikové sk.	n_{1j}	n_{2j}	n_j

Tabulka 3: Kontingenční tabulka pro dvě porovnávané skupiny v čase přežití $t_{(j)}$

V dané tabulce označuje n_{1j} počet jedinců v riziku v čase $t_{(j)}$ v rámci první skupiny a n_{2j} počet jedinců v riziku v čase $t_{(j)}$ v rámci druhé skupiny. Dále d_{1j} označuje počet událostí (např. úmrtí, pooperačních komplikací) v čase $t_{(j)}$ v první skupině a d_{2j} počet událostí v čase $t_{(j)}$ v druhé skupině. Celkový počet jedinců v riziku v čase $t_{(j)}$ je $n_j = n_{1j} + n_{2j}$ a celkový počet událostí v čase $t_{(j)}$ je $d_j = d_{1j} + d_{2j}$.

Za předpokladu nulové hypotézy, že funkce přežití jsou v obou dvou skupinách stejné, můžeme d_{1j} považovat za náhodnou proměnnou. Ve skutečnosti d_{1j} má *hypergeometrické rozdělení*, na základě něhož můžeme spočítat pravděpodobnost, že náhodná proměnná

bude mít hodnotu právě d_{1j} jako

$$\frac{\begin{pmatrix} d_j \\ d_{1j} \end{pmatrix}\begin{pmatrix} n_j - d_j \\ n_{1j} - d_{1j} \end{pmatrix}}{\begin{pmatrix} n_j \\ n_{1j} \end{pmatrix}}$$

Dále střední hodnota hypergeometrické náhodné proměnné d_{1j} je rovna

$$e_{1j} = \frac{n_{1j} d_j}{n_j}.$$

Tuto hodnotu označujeme jako **očekávaný** počet událostí v čase $t_{(j)}$.

Nyní spočítáme rozdíl mezi očekávanou hodnotou a skutečnou v každém čase $t_{(j)}$. Získáme tak statistiku

$$U_L = \sum_{j=1}^{m} (d_{1j} - e_{1j})$$

Tato statistika bude mít nulovou střední hodnotu, pokud $E(d_{1j}) = e_{1j}$. Navíc jestliže časy událostí jsou nezávislé, pak rozptyl statistiky U_L je roven součtu rozptylů d_{1j}. A protože d_{1j} se řídí hypergeometrickým rozdělením, je rozptyl d_{1j} roven

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Takže rozptyl U_L je

$$DU_L = \sum_{j=1}^{m} v_{1j} = V_L.$$

Lze ukázat, že pokud máme dostatek pozorovaných časů událostí, můžeme U_L aproximovat normálním rozdělením. Tedy

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1), \qquad \text{nebo} \qquad \frac{U_L^2}{V_L} \sim \chi_1^2.$$

Tento postup byl odvozen Mantelem a Haenszelem (1959) a je znám jako **Mantel-Haenszelova** procedura. Ve skutečnosti má tento test různé názvy např. *Mantel-Cox*, či *Peto-Mantel-Haenszel*, ale nejvíc se pro něj vžil termín *Log-rank test*. Kromě tohoto testu lze použít ke srovnání dvou skupin **Wilcoxonův test**, který je také známý jako **Breslowův test**, dále existuje **Taron-Wareův** test nebo **Peto-Prenticeův** test. Více k nim lze nalézt v [3], [10].

4 Coxův proporcionální hazardní model

$$h(t) = h_0(t)e^{\beta' x} \tag{4.1}$$

4.1 Odhad parametrů PH modelu

Při řešení problému, jak odhadnout parametry modelu (4.1), se většinou nejprve zamyslíme, jestli bychom nemohli využít metody maximální věrohodnosti MLE.

Předpokládejme, že máme n nezávislých pozorování, kde každé obsahuje informaci o době do pooperačních komplikací, jestli se jedná o čas cenzorovaný či nikoli a proměnné, které byly určeny na začátku studie a v průběhu této studie se nemění. [pozn. V praxi je opravdu běžné, že se hodnoty naměří na začátku a dále zůstávají stejné. Přesto může nastat situace, že se proměnné v průběhu času měnit budou. Takové proměnné mohou být jednoduše přizpůsobeny. Více k tomuto problému se lze dočíst v [10] Kapitola VII.]

Bez újmy na obecnosti budeme pro jednoduchost pracovat pouze s jednou vysvětlující proměnnou. Označme si trojici pozorovaného času, proměnné a informaci o cenzorování $(t_i, x_i, c_i), i = 1, 2, \ldots, n$.

Prvním krokem je sestavení věrohodnostní funkce. S tou jsme se již setkali a řekli jsme, že na ni můžeme pohlížet jako na pravděpodobnost toho, že naměříme opět stejná data.

Příklad 4.1
Doba do pooperačních komplikací je spojitá náhodná veličina pocházející z rozdělení s hustotou pravděpodobnosti $f(t, \beta, x)$. Pokud jsme naměřili hodnoty $(10, 45, 1)$, $(11, 40, 0)$, $(15, 39, 0)$ a $(18, 25, 1)$, pak věrohodnostní funkce bude dána jako

$$\ell(\beta) = f(10, \beta, 45) \cdot S(11, \beta, 40) \cdot S(15, \beta, 39) \cdot f(18, \beta, 25).$$

■

Z příkladu (4.1) vidíme, že obecně můžeme věrohodnostní funkci zapsat

$$\ell(\beta) = \prod_{i=1}^{n} \left\{ [f(t_i, \beta, x_i)]^{c_i} \cdot [S(t_i, \beta, x_i)]^{1-c_i} \right\}. \tag{4.2}$$

Lépe se nám bude pracovat s logaritmem věrohodnostní funkce

$$L(\beta) = \sum_{i=1}^{n} \left\{ c_i ln\left[f(t_i, \beta, x_i)\right] + (1 - c_i)\, ln\left[S(t_i, \beta, x_i)\right] \right\}. \tag{4.3}$$

Postup nalezení MLE odhadu spočívá v derivování $L(\beta)$ podle neznámého parametru β, položení této derivace rovno nule a vyřešení pro β.

Ze vztahu (2.1) si můžeme vyjádřit hustotu pravděpodobnosti $f(t, x, \beta)$ jako součin hazardní funkce $h(t, x, \beta)$ a funkce přežití $S(t, x, \beta)$. Tento vztah můžeme dosadit do (4.3) a využitím (3.10) a (3.12) po úpravách dostaneme

$$L(\beta) = \sum_{i=1}^{n} \left\{ c_i ln\left[h_0(t_i)\right] + c_i x_i \beta + e^{x_i \beta} ln\left[S_0(t_i)\right] \right\}. \tag{4.4}$$

Plná maximalizace věrohodnostní funkce vyžaduje, abychom maximalizovali výraz (4.4) s ohledem na parametr β, který nás zajímá, na blíže nespecifikované základní hazardní funkci a funkci přežití. Kalbfleisch a Prentice [12] ukázali, že k tomu není možné použít log-věrohodnostní funkci (4.4).

Cox (1972) navrhl vztah, který nazýváme **částečnou věrohodnostní funkcí**. Tato funkce závisí pouze na parametrech β, které nás zajímají. Předpokládal, že výsledné odhady parametrů z částečné věrohodnostní funkce budou mít stejné rozdělení pravděpodobnosti jako plně maximalizované věrohodnostní odhady. Důkaz této domněnky lze najít v [10], resp. nástin myšlenky nalezneme v [26].

Částečná věrohodnostní funkce je dána vztahem

$$\ell_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_i\beta}}{\sum\limits_{j \in R(t_i)} e^{x_j\beta}} \right]^{c_i}, \tag{4.5}$$

kde sumace ve jmenovateli jde přes všechny jednotlivce, kteří jsou v čase t_i v rizikové skupině, kterou označujeme jako $R(t_i)$.

Je třeba zdůraznit, že vztah (4.5) je založen na předpokladu, že se v datech neopakují pozorované časy. Výraz (4.5) lze dále upravit tak, abychom vyloučili členy u kterých $c_i = 0$.

$$\ell_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta}}{\sum\limits_{j \in R(t_{(i)})} e^{x_j\beta}} \tag{4.6}$$

Log-částečná věrohodnostní funkce je

$$L_p(\beta) = \sum_{i=1}^{m} \left\{ x_{(i)}\beta - ln \left[\sum_{j \in R(t_{(i)})} e^{x_j\beta} \right] \right\} \tag{4.7}$$

Hledáme maximum této funkce, proto ji zderivujeme podle β a položíme rovno nule. Takto nalezený odhad značíme $\hat{\beta}$.

$$\begin{aligned} \frac{\partial L_p(\beta)}{\partial \beta} &= \sum_{i=1}^{m} \left\{ x_{(i)} - \frac{\sum\limits_{j \in R(t_{(i)})} x_j e^{x_j\beta}}{\sum\limits_{j \in R(t_{(i)})} e^{x_j\beta}} \right\} \\ &= \sum_{i=1}^{m} \left\{ x_{(i)} - \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_j \right\} \\ &= \sum_{i=1}^{m} \left\{ x_{(i)} - \bar{x}_{w_i} \right\}, \end{aligned} \tag{4.8}$$

kde

$$w_{ij} = \frac{e^{x_j\beta}}{\sum\limits_{l \in R(t_{(i)})} e^{x_l\beta}}, \qquad \bar{x}_{w_i} = \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_j.$$

Odhad rozptylu a **standardní chyby** odhadu koeficientů jsou dány vztahy

$$\hat{D}\hat{\beta} = I(\hat{\beta})^{-1}, \qquad \hat{SE}(\hat{\beta}) = \sqrt{\hat{D}\hat{\beta}} \tag{4.9}$$

kde $I(\beta)$ se nazývá **pozorovaná informace** a je dána vztahem

$$I(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta^2}, \tag{4.10}$$

kde druhou derivaci z částečné věrohodnostní funkce můžeme zkráceně po úpravách zapsat

$$\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} w_{ij}(\beta)(x_j - \bar{x}_{w_i})^2. \tag{4.11}$$

V případě, že máme více než jednu proměnnou např. p vysvětlujících proměnných, hovoříme o $I(\hat{\beta})$ jako o **pozorované informační matici** jejíž prvky jsou druhé derivace

$$I_{i,j}(\beta) = \frac{\partial^2 L(\beta)}{\partial \beta_i \beta_j}, \quad i = 1, \ldots, p; \quad j = 1, \ldots, p.$$

Odmocniny diagonálních prvků inverze této matice jsou standardní chyby SE odhadovaných parametrů $\beta_1, \beta_2, \ldots, \beta_p$.

4.2 Odvození částečné věrohodnostní funkce

Vraťme se k částečné věrohodnostní funkci (4.5) a pokusme se ji odvodit [3].

Již dříve jsem se zmínila, co je to věrohodnostní funkce, a jak vypadá pro náš Coxův model. Základní argument používaný při konstrukci věrohodnostní funkce proporcionálního hazardního modelu je, že intervaly mezi po sobě jdoucími událostmi (pooperační komplikace) neposkytují informaci o vlivu vysvětlujících proměnných na riziko pooperačních komplikací. Je to protože základní hazardní funkce má libovolný tvar, a tak je možné, že $h_0(t)$ (a proto i $h(t)$) je v těchto časových intervalech, kde se nevyskytují pooperační komplikace, nulová. Tím se zkrátka myslí, že tyto intervaly neposkytují žádné informace o hodnotách parametrů β. Proto uvažujeme pravděpodobnost, že *i*-tý jedinec bude mít pooperační komplikace v některém čase $t_{(j)}$ za podmínky, že $t_{(j)}$ je jedno z pozorovaných časů pooperačních komplikací $t_{(1)}, t_{(2)}, \ldots, t_{(m)}$. Pokud u jedince, který měl v čase $t_{(j)}$ pooperační komplikace, jsou vysvětlující proměnné dány vektorem $x_{(j)}$, pak je tato pravděpodobnost

$$P(\text{jedinec s prom. } x_{(j)} \text{ bude mít komplikace v } t_{(j)} | \text{v } t_{(j)} \text{ došlo ke komplikacím}). \tag{4.12}$$

Nyní můžeme využít vztahu

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Protože v tomto případě $A \subset B$, pravděpodobnost ze vztahu (4.12) přejde do tvaru

$$\frac{P(\text{jedinec s proměnnými } x_{(j)} \text{ bude mít pooperační komplikace v čase } t_{(j)})}{P(\text{v čase } t_{(j)} \text{ došlo k pooperačním komplikacím})}. \tag{4.13}$$

A protože předpokládáme, že časy pooperačních komplikací jsou na sobě nezávislé, můžeme jmenovatele ve vztahu(4.13) napsat jako sumu pravděpodobností pooperačních komplikací v čase $t_{(j)}$ přes všechny jedince, kteří jsou v tomto čase v rizikové skupině $R(t_{(j)})$.

$$\frac{P(\text{jedinec s proměnnými } x_{(j)} \text{ bude mít pooperační komplikace v čase } t_{(j)})}{\sum\limits_{l \in R(t_{(j)})} P(\text{jedinec l bude mít pooperační komplikace v čase } t_{(j)})}. \tag{4.14}$$

Abychom mohli přejít od pravděpodobnosti k hazardní funkci, pravděpodobnost pooperačních komplikací v čase $t_{(j)}$ ve vztahu (4.14) nahradíme pravděpodobností pooperačních komplikací v intervalu $(t_{(j)}, t_{(j)} + \delta t)$ a podělíme čitatele i jmenovatele δt. Dostaneme

$$\frac{P(\text{jedinec s proměnnými } x_{(j)} \text{ bude mít pooperační komplikace v čase } (t_{(j)}, t_{(j)} + \delta t))/\delta t}{\sum\limits_{l \in R(t_{(j)})} P(\text{jedinec l bude mít pooperační komplikace v čase } (t_{(j)}, t_{(j)} + \delta t))/\delta t}.$$

Půjdeme-li v tomto výrazu limitně s $\delta t \to 0$, dostaneme tak vlastně podíl (4.14), ale zároveň víme, že takto definovaná limita odpovídá riziku, že v čase $t_{(j)}$ dojde k pooperačním komplikacím (hazardní funkce). Takže můžeme psát

$$\frac{\text{Riziko pooperační komplikace v čase } t_{(j)} \text{ u jedince s proměnnými } x_{(j)}}{\sum\limits_{l \in R(t_{(j)})} (\text{Riziko pooperační komplikace v čase } t_{(j)} \text{ u jedince l})} \tag{4.15}$$

Pokud v čase $t_{(j)}$ bude mít pooperační komplikace *i*-tý jedinec, můžeme hazardní funkci v čitateli zapsat jako $h_i(t_{(j)})$. Podobně ve jmenovateli jako $h_l(t_{(j)})$. Tak přejdeme od výrazu (4.12)k výrazu

$$\frac{h_i(t_{(j)})}{\sum\limits_{l \in R(t_{(j)})} h_l(t_{(j)})} \tag{4.16}$$

Do vztahu (4.16) dosadíme (3.10). Základní hazardní funkce $h_0(t)$ se pokrátí a dostáváme

$$\frac{e^{(\beta' x_{(j)})}}{\sum\limits_{l \in R(t_{(j)})} e^{(\beta' x_{(l)})}} \tag{4.17}$$

A nakonec součin (4.17) přes všechny pozorované časy m nám dá částečnou věrohodnostní funkci (4.5).

4.3 Opakující se pozorované časy v datech

Částečnou věrohodnostní funkci (4.5) jsme založili na předpokladu, že se v datech nevyskytují opakovaná pozorování času událostí. V případě, že chceme podchytit tyto tzv. *tied data*, musíme částečnou věrohodnostní funkci modifikovat. Popíšeme si čtyři metody, které tuto modifikaci realizují.

Pro usnadnění vyjdeme u všech zmíněných metod z následujícího konkrétního příkladu pozorovaných dat:

i	x_i	c_i	t_i
1	5	1	2
2	4	1	2
3	6	0	3
4	2	1	4
5	2	1	5

Vidíme, že v prvním a druhém případě máme stejná pozorování času událostí. V datech se tedy vyskytují tři různá pozorování času událostí (t_1, t_4, t_5). Částečnou věrohodnostní funkci můžeme proto zapsat jako

$$\ell(\beta) = \prod_{j=1}^{3} \ell_j(\beta) = \ell_1(\beta)\ell_2(\beta)\ell_3(\beta).$$

Protože se časy t_4 a t_5 v datech neopakuji, můžeme komponenty částečné věrohodnostní funkce $\ell_2(\beta)$ a $\ell_3(\beta)$ vypočítat standardním způsobem

$$\ell_2(\beta) = \frac{e^{x_4\beta}}{e^{x_4\beta} + e^{x_5\beta}}, \quad a \quad \ell_3(\beta) = 1.$$

Co nás tedy bude ve skutečnosti zajímat je komponenta $\ell_1(\beta)$.

Přesná metoda

Je založena na předpokladu, že čas události je spojitá náhodná veličina a že časy t_1 a t_2 jsou ve skutečnosti různé. Vlivem zaokrouhlovacích chyb (např. na dny, měsíce, roky) se nám tak určitá informace o času události ztratila.

Protože ale nevíme, který čas předcházel jakému, uvažujeme všechny možnosti, které by mohly nastat. V našem případě to je $2! = 2$. Označme A_1 případ, ve kterém $t_1 < t_2$ a A_2 případ, ve kterém $t_2 < t_1$. Pak

$$\ell_1(\beta) = P(\text{v čase } t_1 \text{ dojde ke dvěma úmrtím}) = P(A_1 \cup A_2) = P(A_1) + P(A_2),$$

kde $P(A_1)$ a $P(A_2)$ spočítáme obvyklým způsobem:

$$\begin{aligned} P(A_1) &= \frac{e^{x_1\beta}}{e^{x_1\beta} + e^{x_2\beta} + e^{x_3\beta} + e^{x_4\beta} + e^{x_5\beta}} \cdot \frac{e^{x_2\beta}}{e^{x_2\beta} + e^{x_3\beta} + e^{x_4\beta} + e^{x_5\beta}} \\ P(A_2) &= \frac{e^{x_2\beta}}{e^{x_2\beta} + e^{x_1\beta} + e^{x_3\beta} + e^{x_4\beta} + e^{x_5\beta}} \cdot \frac{e^{x_1\beta}}{e^{x_1\beta} + e^{x_3\beta} + e^{x_4\beta} + e^{x_5\beta}} \end{aligned}$$

Tato metoda může být bohužel výpočetně velmi náročná, a proto se ve softwarových balíčcích často využívá spíše aproximace přesné částečné věrohodnostní funkce.

[27]

Breslowova aproximace

U našeho příkladu můžeme provést aproximaci následovně

$$\begin{aligned}\frac{e^{x_2\beta}}{e^{x_2\beta}+e^{x_3\beta}+e^{x_4\beta}+e^{x_5\beta}} &\approx \frac{e^{x_2\beta}}{e^{x_1\beta}+e^{x_2\beta}+e^{x_3\beta}+e^{x_4\beta}+e^{x_5\beta}}\\ \frac{e^{x_1\beta}}{e^{x_1\beta}+e^{x_3\beta}+e^{x_4\beta}+e^{x_5\beta}} &\approx \frac{e^{x_1\beta}}{e^{x_1\beta}+e^{x_2\beta}+e^{x_3\beta}+e^{x_4\beta}+e^{x_5\beta}}\end{aligned}$$

Pak $P(A_1)$ i $P(A_2)$ a tudíž i $\ell_1(\beta)$ může být aproximováno jako

$$\frac{e^{x_1\beta}}{e^{x_1\beta}+e^{x_2\beta}+e^{x_3\beta}+e^{x_4\beta}+e^{x_5\beta}}\cdot\frac{e^{x_2\beta}}{e^{x_1\beta}+e^{x_2\beta}+e^{x_3\beta}+e^{x_4\beta}+e^{x_5\beta}}=\frac{e^{(x_1+x_2)\beta}}{\left[\sum_{l=1}^{5}e^{x_l\beta}\right]^2}$$

Jestliže v $j-$tém odlišném pozorovaném čase nastane d_j stejných pozorovaných časů, pak obecný zápis $\ell_j(\beta)$ komponenty je

$$\ell_j(\beta)\approx\frac{exp\left(\beta\sum\limits_{l\in D(t_{(j)})}x_l\right)}{\left[\sum\limits_{l\in R(t_{(j)})}exp(x_l\beta)\right]^{d_j}},$$

kde $R(t_{(j)})$ je riziková skupina v čase $t_{(j)}$ a $D(t_{(j)})$ je skupina jedinců, jejichž časy přežití jsou rovny $t_{(j)}$. Pro částečnou věrohodnostní funkci platí

$$\ell(\beta)=\prod_{j=1}^{m}\ell_j(\beta)\approx\prod_{j=1}^{m}\frac{exp\left(\beta\sum\limits_{l\in D(t_{(j)})}x_l\right)}{\left[\sum\limits_{l\in R(t_{(j)})}exp(x_l\beta)\right]^{d_j}},\tag{4.18}$$

kde m je celkový počet odlišných pozorovaných času přežití. Aproximaci (4.18) navrhl Breslow (1974).

Pokud bude d_j malé a počet jedinců v rizikové skupině n_j bude velké (tedy podíl d_j/n_j bude malý), pak bude Breslowova aproximace fungovat dobře (aproximace bude blízká přesné částečné věrohodnostní funkci). Pokud ale tyto podmínky nebudou splněny, pak tato aproximace nebude příliš vhodná. Proto Efron (1977) navrhl jiný postup.

[27]

Efronova aproximace

Co se týká našeho příkladu, můžeme přesnou částečnou věrohodnostní funkci zapsat schematicky

$$\ell_1(\beta) = \frac{bc}{a(a-b)} + \frac{bc}{a(a-c)},$$

což můžeme aproximovat vztahem

$$\ell_1(\beta) \approx \frac{2bc}{a(a-(b+c)/2)}.$$

Tato aproximace je základní myšlenkou Efronovy aproximace

$$\ell(\beta) \approx \prod_{j=1}^{m} \frac{exp\left(\beta \sum\limits_{l \in D(t_{(j)})} x_l\right)}{\prod\limits_{k=1}^{d_j} \left(\sum\limits_{l \in R(t_{(j)})} e^{x_l\beta} - \frac{k-1}{d_j} \sum\limits_{l \in D(t_{(j)})} e^{x_l\beta}\right)} \tag{4.19}$$

[27]

Coxova aproximace

Tato aproximace není v softwarových balíčcích tak běžná, přesto jsem se rozhodla ji sem zařadit. Byla navržena Coxem (1972).

$$\ell(\beta) \approx \prod_{j=1}^{m} \frac{exp\left(\beta \sum\limits_{l \in D(t_{(j)})} x_l\right)}{\sum\limits_{\text{all } D(t_{(j)})} exp\left(\beta \sum\limits_{k \in D(t_{(j)})} x_k\right)} \tag{4.20}$$

Vztah (4.20) je založen na modelu, který odpovídá situaci s diskrétní časovou škálou. [pozn. Jsou tedy přípustné opakování pozorovaných časů v datech.] Diskrétní tvar proporcionálního hazardního modelu (3.10) je

$$\frac{h(t)}{1-h(t)} = \frac{h_0(t)}{1-h_0(t)} e^{x\beta},$$

jehož věrohodnostní funkce je dána právě vztahem (4.20). Pokud velikost diskrétního časového intervalu půjde limitně k nule, bude se blížit k proporcionálnímu modelu (3.10).

Pokud v datech nejsou žádné opakující se pozorované časy, tedy $d_j = 1$ pro $j = 1, 2, \ldots, m$, pak aproximace (4.18), (4.19) a (4.20) odpovídají věrohodnostní funkci (4.6).

[3]

4.4 Newton-Raphsonova metoda

K maximalizaci částečné věrohodnostní funkce se často používá iterační Newton-Rhapsonova metoda.

Nechť β označuje vektor neznámých parametrů, tedy $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ a $u(\beta)$ je $p \times 1$ vektor prvních derivací logaritmu částečné věrohodnostní funkce

$$u(\beta) = \left(\frac{\partial L(\beta)}{\partial \beta_1}, \frac{\partial L(\beta)}{\partial \beta_2}, \ldots, \frac{\partial L(\beta)}{\partial \beta_p} \right)^T.$$

Dále označme $I(\beta)$ informační matici $p \times p$, o které jsem se již zmínila dříve. Zopakuji, že (i, j)-tá složka matice $I(\beta)$ je

$$- \frac{\partial^2 L(\beta)}{\partial \beta_i \partial \beta_j}.$$

Odhad vektoru parametrů β je dán iteračním vztahem

$$\hat{\beta}_{k+1} = \hat{\beta}_k + I^{-1}(\hat{\beta}_k) u(\hat{\beta}_k) \tag{4.21}$$

Metodu můžeme zahájit s aproximací $\hat{\beta}_0 = 0$. Proces ukončíme pokud změna $L(\beta)$ bude malá nebo pokud maximum z absolutního rozdílu $|\hat{\beta}_{k+1} - \hat{\beta}_k|$ bude malé.

Pokud bude iterační metoda konvergovat, může být kovarianční matice odhadovaných parametrů aproximována inverzní maticí $I(\hat{\beta})^{-1}$. Odmocnina z diagonály této matice je pak standardní chybou SE odhadovaných hodnot parametrů $\beta_1, \beta_2, \ldots, \beta_p$. [3]

4.5 Intervaly spolehlivosti a testy hypotéz pro parametry β

K určení aproximací intervalů spolehlivosti neznámých parametrů $\beta_1, \beta_2, \ldots, \beta_p$ využíváme standardní chyby odhadnutých parametrů. Interval spolehlivosti pro k-tý parametr je dán vztahem

$$\left\langle \hat{\beta}_k - z_{\alpha/2} SE(\hat{\beta}_k); \hat{\beta}_k + z_{\alpha/2} SE(\hat{\beta}_k) \right\rangle,$$

kde $\hat{\beta}$ je odhad parametru β a $z_{\alpha/2}$ je $\frac{\alpha}{2}$ kvantil normovaného normálního rozdělení.

Příklad 4.2
Máme-li určit 95%-ní intervaly spolehlivosti pro parametry β_1 a β_2, když $\hat{\beta}_1 = -0.659$, $\hat{\beta}_2 = 0.097$, $SE(\hat{\beta}_1) = 0.2153$, $SE(\hat{\beta}_2) = 0.2841$, pak dostaneme dle vztahu

$$\left\langle \hat{\beta}_k - z_{0.025} SE(\hat{\beta}_k); \hat{\beta}_k + z_{0.025} SE(\hat{\beta}_k) \right\rangle$$

$\hat{\beta}_1 : \langle -0.659 - 1.96 \cdot 0.2153; -0.659 + 1.96 \cdot 0.2153 \rangle = \langle -1.081; -0.237 \rangle$

$\hat{\beta}_2 : \langle 0.097 - 1.96 \cdot 0.2841; 0.097 + 1.96 \cdot 0.2841 \rangle = \langle -0.460; 0.654 \rangle$

Pokud interval neobsahuje nulu, je to signál, že takový parametr nebude nulový (bude zahrnut v modelu). ■

Co se týká testování hypotéz k posouzení významnosti parametrů β, existují tři běžně používané testy.

Test poměru částečné věrohodnostní funkce

Tento test, označovaný písmenem G, je dán vztahem

$$G = 2\left\{L(\hat{\beta}) - L(0)\right\}, \tag{4.22}$$

kde $\hat{\beta}$ je vektor odhadnutých parametrů $(\hat{\beta}_1, \ldots, \hat{\beta}_p)$ a $L(0)$ je log-částečná věrohodnostní funkce s nulovým vektorem parametrů β, tedy

$$L(0) = -\sum_{i=1}^{m} ln(n_i),$$

kde n_i označuje počet jedinců v rizikové skupině v čase $t_{(i)}$.

Statistika G se řídí χ^2 rozdělením s p stupni volnosti (za každý odhadovaný parametr jeden stupeň volnosti). Testujeme nulovou hypotézu

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_A : Aspoň jeden z koeficientů je různý od nuly.

Potom p-hodnota pro tento test je dána jako

$$P\left(\chi^2(p) \geq 2\left\{L(\hat{\beta}) - L(0)\right\}\right).$$

Waldova statistika

Nechť β je vektor parametrů $(\beta_1, \ldots, \beta_p)$. Waldovu statistiku můžeme spočítat dvěma způsoby. Ve statistických programech se běžně používá postup, kdy se zvlášť pro každý odhadovaný parametr β_k stanoví Waldova statistika. Testuje se hypotéza

$H_0 : \beta_k = 0$ v případě, že jsou v modelu obsaženy všechny ostatní parametry.
$H_A : \beta_k \neq 0$ v případě, že jsou v modelu obsaženy všechny ostatní parametry.

Testujeme na základě vypočtené hodnoty statistiky

$$z = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)},$$

která se řídí normovaným normálním rozdělením. p-hodnota je pak rovna $P(|z| > \hat{\beta}_k/SE(\hat{\beta}_k))$. V programech, jako je např. SPSS, se statistika z modifikuje tak, že se umocní na druhou. Tato upravená statistika se pak řídí χ^2 rozdělením s 1 stupněm volnosti.

V druhém případě si označíme vektor prvních derivací log-částečně věrohodnostní funkce jako $u(\beta)$. Na základě hypotézy H_0

vektor $u(0) = u(\beta)|_{\beta=0}$ se řídí multivariačním normálním rozdělením s vektorem středních hodnot roven nule a kovarianční maticí danou informační maticí $I(0) = I(\beta)|_{\beta=0}$. Waldova statistika je pak dána vztahem

$$\hat{\beta}' I(\hat{\beta})\hat{\beta},$$

a asymptoticky se řídí χ^2 rozdělením s p stupni volnosti.

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_A : Aspoň jeden z parametrů je různý od nuly.

Skóre test

Posledním testem je skóre test, na který lze také nahlížet dvěma způsoby. V prvním případě se pro každý parametr β_k stanoví skóre zvlášť (stejně jako u Waldovy statistiky).

$H_0 : \beta_k = 0$ v případě, že jsou v modelu obsaženy všechny ostatní parametry.
$H_A : \beta_k \neq 0$ v případě, že jsou v modelu obsaženy všechny ostatní parametry.

Testovanou statistikou je

$$z^* = \left. \frac{\partial L(\beta_k)/\partial \beta_k}{\sqrt{I(\beta_k)}} \right|_{\beta_k = 0},$$

která se řídí normovaným normálním rozdělením. Opět lze statistiku umocnit na druhou. Pak se bude řídit obdobně jako u Waldovy statistiky χ^2 rozdělením s 1 stupněm volnosti.

V druhém případě bude hypotéza opět stejná jako u Waldovy statistiky.

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_A : Aspoň jeden z parametrů je různý od nuly.

Statistika skóre testu je pak dána vztahem

$$u'(0)\,[I(0)]^{-1}\,u(0) \tag{4.23}$$

a řídí se χ^2 rozdělením s p stupni volnosti.

4.6 Odhad hazardní funkce a funkce přežití

Předpokládejme, že proporcionální hazardní model obsahuje p vysvětlujících proměnných $X_1, X_2, \ldots, X_p$. Příslušné odhady koeficientů těchto proměnných jsou $\beta_1, \beta_2, \ldots, \beta_p$. Odhadovaná hazardní funkce i-tého jedince z celkového počtu n jedinců je dána vztahem

$$\hat{h}_i(t) = exp(\hat{\beta}' x_i)\hat{h}_0(t), \tag{4.24}$$

kde x_i je vektor vysvětlujících proměnných i-tého jedince, $i = 1, 2, \ldots, n$, $\hat{\beta}$ je vektor odhadovaných koeficientů a $\hat{h}_0(t)$ je odhad základní hazardní funkce. Hazardní funkci jedince můžeme tedy odhadnout, jakmile najdeme odhad základní hazardní funkce.

Odhad základní hazardní funkce byl odvozen **Kalbfeischem** a **Prenticem** (1973). Předpokládejme, že máme m seřazených odlišných necenzorovaných pozorovaných časů úmrtí $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$, dále d_j označuje počet úmrtí v čase $t_{(j)}$ a n_j označuje počet jedinců v rizikové skupině v čase $t_{(j)}$. Odhad základní hazardní funkce v čase $t_{(j)}$ je dán vztahem

$$\hat{h}_0(t) = 1 - \hat{\xi}_j, \tag{4.25}$$

kde $\hat{\xi}_j$ je řešením rovnice

$$\sum_{l \in D(t_{(j)})} \frac{exp(\hat{\beta}' x_l)}{1 - \hat{\xi}_j^{exp(\hat{\beta}' x_l)}} = \sum_{l \in R(t_{(j)})} exp(\hat{\beta}' x_l) \tag{4.26}$$

pro $j = 1, 2, \ldots, m$. V rovnici (4.26) označuje $D(t_{(j)})$ skupinu d_j jedinců, kteří zemřeli v čase $t_{(j)}$ a $R(t_{(j)})$ označuje skupinu n_j jedinců, kteří jsou v čase $t_{(j)}$ v rizikové skupině.

V konkrétním případě, ve kterém se nevyskytují opakovaná pozorování času úmrtí (nejsou v datech přítomny tzv. *tied data* a tedy $d_j = 1$ pro všechna $j = 1, 2, \ldots, m$), suma na levé straně rovnice (4.26) bude obsahovat pouze jeden člen a $\hat{\xi}_j$ se spočte jako

$$\hat{\xi}_j = \left(1 - \frac{exp(\hat{\beta}' x_{(j)})}{\sum\limits_{l \in R(t_{(t)})} exp(\hat{\beta}' x_l)} \right)^{exp(\hat{\beta}' x_{(j)})}, \tag{4.27}$$

kde $x_{(j)}$ je vektor vysvětlujících proměnných jedince, který zemřel v čase $t_{(j)}$.

V případě, že se v datech nějaké pozorované časy opakují, nelze vyřešit rovnici (4.26) explicitně a je třeba využít některých iteračních postupů.

Předpokládejme, že hazardní funkce je mezi sousedními pozorovanými časy úmrtí konstantní. Vhodný odhad základní hazardní funkce v takovémto intervalu pak získáme, pokud podělíme (4.25) časovým intervalem. Dostaneme

$$\hat{h}_0(t) = \frac{1 - \hat{\xi}_j}{t_{(j+1)} - t_{(j)}}, \tag{4.28}$$

kde $t_{(j)} \leq t < t_{(j+1)}$, $j = 1, 2, \ldots, m-1$ a $\hat{h}_0(t) = 0$ pro $t < t_{(1)}$.

Na $\hat{\xi}_j$ můžeme nahlížet jako na pravděpodobnost, že jedinec přežije přes interval $\langle t_{(j)}, t_{(j+1)} \rangle$. Základní funkci přežití lze pak tedy odhadnout jako

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\xi}_j, \tag{4.29}$$

pro $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \ldots, m-1$. Odhad základní funkce přežití je roven jedné pokud $t < t_{(1)}$ a nule pokud $t \geq t_{(m)}$ za předpokladu, že se v datech nevyskytují cenzorované časy, které by byly větší než $t_{(m)}$. Jinak se za $\hat{S}_0(t)$ vezme hodnota $\hat{S}_0(t_{(m)})$, a to až do největšího cenzorovaného času, avšak odhad funkce přežití není po tomto čase již definován.

Základní kumulativní hazardní funkce se pak získá jako

$$\hat{H}_0(t) = -ln\hat{S}_0(t) = -\sum_{j=1}^{k} ln\hat{\xi}_j, \tag{4.30}$$

pro $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \ldots, m-1$ a $\hat{H}_0(t) = 0$ pro $t < t_{(1)}$.

Na základě odhadů základní hazardní funkce, základní funkce přežití a základní kumulativní hazardní funkce v rovnicích (4.28), (4.29), a (4.30) můžeme odhadnout hazardní funkci i-tého jedince s vektorem vysvětlujících proměnných x_i, a to díky vztahu (4.24). Pokud obě strany tohoto vztahu zintegrujeme, získáme vztah pro kumulativní hazardní funkci i-tého jedince [pozn. dle vztahu (2.2)]

$$\hat{H}_i(t) = exp(\hat{\beta}' x_i)\hat{H}_0(t). \tag{4.31}$$

Pokud levou i pravou stranu rovnice (4.31) vynásobíme -1 a použijeme exponencionální funkci, dostaneme vztah odhadu funkce přežití i-tého jedince

$$\hat{S}_i(t) = \left\{\hat{S}_0(t)\right\}^{exp(\hat{\beta}' x_i)}, \tag{4.32}$$

pro $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \ldots, m-1$.

V případě, že v datech nejsou žádné vysvětlující proměnné (máme jen časy přežití), rovnice (4.26) přejde do tvaru

$$\frac{d_j}{1 - \hat{\xi}_j} = n_j,$$

ze kterého vyjádříme $\hat{\xi}_j$

$$\hat{\xi}_j = \frac{n_j - d_j}{n_j}$$

Pak odhad základní hazardní funkce v čase $t_{(j)}$ je $1 - \hat{\xi}_j$, což je d_j/n_j. A odhad funkce přežití po dosazení vztahu (4.29) do vztahu (4.32) a následné nahrazení $\hat{\xi}_j$ výše uvedeným vztahem nám dá

$$\hat{S}_i(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right),$$

což odpovídá již dříve zmíněnému Kaplan-Meierovu odhadu funkce přežití. Ukázali jsme, že odhad funkce přežití ve vztahu (4.32) je zobecněním Kaplan-Meierova odhadu pro případy, kdy hazardní funkce závisí na vysvětlujících proměnných.

Jak už jsem se zmínila, v případě, že se v datech vyskytují opakovaná pozorování času přežití, odhad základní hazardní funkce lze spočítat pomocí iteračních metod. Tomuto postupu se můžeme vyhnout, pokud využijeme aproximace sumy na levé straně rovnice (4.26). Tyto aproximace lze najít v [3] v kapitole *3.8.2*.

5 Analýza dat

Další část této práce se zabývá aplikací nastudovaných metod na konkrétní data. Rozebereme, s jakými daty budeme pracovat, sestavíme Coxův proporcionální model a vytvoříme jednoduchý program, jehož úkolem bude vybrat vhodné vysvětlující proměnné, navrhnout model a odhadnout základní hazardní funkci.

5.1 Seznámení s daty

Díky spolupráci s FNO máme k dispozici data zaznamenávající dobu do pooperačních komplikací pacientů po operaci kolorekta. U každého pacienta byly sledovány následující proměnné:

- doba do pooperačních komplikací (*uvedeno ve dnech*)
- cenzorovaný údaj (*0-cenzorovaný, 1-necenzorovaný*)
- skupina (*odpovídá druhu operace, 0-otevřená, 1-laparoskopická*)
- věk
- pohlaví (*0-muž, 1-žena*)
- diagnóza (*0-c18, 1-c19, 2-c20*)
- BMI (*Body Mass Index*)
- ASA (*American Society of Anaesthesiology Classification, kódováno hodnotami 1-4*)
- ICHS (*0-ne, 1-ano*)
- arytmie (*0-ne, 1-ano*)
- hypertenze (*0-ne, 1-ano*)
- cebrovaskulární (*0-ne, 1-ano*)
- pulmonální (*0-ne, 1-ano*)
- DM (*0-ne, 1-ano*)
- renální (*0-ne, 1-ano*)
- jaterní (*0-ne, 1-ano*)
- předchozí operace (*0-ne, 1-ano*)
- délka operace (*v minutách*)
- perop. komplikace (*0-ne, 1-ano*)

- konverze (*kódováno hodnotami 0-2*)
- krevní ztráta (*v ml*)
- stádium (*kódováno hodnotami 1-4*)
- grading (*kódováno hodnotami 0-2*)

U pacienta je sledováno daleko více proměnných. Z nich jsme vybraly pouze ty, které lékaři určili na základě zkušeností jako možné proměnné, které by mohly ovlivňovat hazardní funkci pacienta.

	n	V procentech
Necenzorované časy $c = 1$	426	54.3%
Cenzorované časy $c = 0$	359	45.7%
Celkem	785	100.0%

Tabulka 4: Četnosti cenzorovaných/necenzorovaných údajů

Z tabulky 4 vidíme, že ve studii je celkem $n = 785$ pacientů, z nichž 426 má necenzorovanou dobu do pooperačních komplikací a 359 cenzorovanou.

5.2 Sestavení modelu

Vidíme, že je zde mnoho potencionálních vysvětlujících proměnných, které by mohly mít vliv na míru rizika u jednotlivých pacientů. Při sestavování modelu je vhodný výběr těchto proměnných jeden z nejdůležitějších kroků. Existuje řada mechanických postupů, jak do modelu vybrat ty správné proměnné, ale ani jeden z nich není dokonalý. Collet v [3] popisuje mechanické postupy **dopředního výběru**, **zpětné eliminace** resp. jejich kombinaci. Zároveň doporučuje držet se obecnější strategie výběru proměnných. Tuto strategii využijeme.

Základem je statistika založená na **testu poměru částečné věrohodnostní funkce** (kapitola 4.5), která se označuje jako $-2ln(\ell(\hat{\beta}))$. Když si uvědomíme, že hodnota částečné věrohodnostní funkce $\ell(\hat{\beta})$ je z intervalu $(0, 1)$, pak poznatek, že čím je $\ell(\hat{\beta})$ větší, tím je model lepší, volně přejde na fakt, že čím je $-2ln\ell(\hat{\beta})$ menší, tím je model lepší. Navíc je zřejmé, že tato hodnota bude vždy kladná.

Pokud chceme určit, zda-li je změna statisticky významná, použijeme princip testu poměru částečné věrohodnostní funkce. Nemusíme ovšem sledovat pouze rozdíl mezi novým modelem a nulovým modelem, jak je tomu v kapitole 4.5, ale můžeme sledovat rozdíl mezi dvěma modely, kde jeden od druhého se liší přidáním resp. odebráním jedné či více vysvětlujících proměnných.

Strategie

1) Pokud máme p vysvětlujících proměnných, sestavíme p modelů, kde každý bude zahrnovat pouze jednu vysvětlující proměnnou. Porovnáme hodnoty $-2ln\ell(\hat{\beta})$ těchto

modelů s hodnotou nulového modelu a určíme, které proměnné významně snižují hodnotu této statistiky.

2) Sestavíme model (označíme ho M_1), který zahrnuje všechny proměnné, které jsme na základě kroku 1) určili jako statisticky významné. Spočítáme příslušnou hodnotu $-2ln\ell(\hat{\beta})$ tohoto modelu. Protože se proměnné v přítomnosti jiných můžou zdát jinak statisticky významné než v případě, kdy je zahrnujeme samostatně, budeme nyní sestavovat modely, které vzniknou z modelu M_1 odebráním vždy jedné proměnné. Ty proměnné, u nichž po jejich odebrání došlo k statisticky významnému nárůstu statistiky $-2ln\ell(\hat{\beta})$, budou v modelu zachovány.
Pokud jsme nějaké proměnné na základě tohoto kroku odebrali, musíme postup znova opakovat. Máme tedy nyní model M_2, který vznikl z M_1 odebráním některých proměnný a na tento model znovu aplikujeme postup z kroku 2). Dokud budeme model redukovat, budeme opakovat i krok 2). Dojdeme k modelu, který si označíme M_n.

3) Proměnné, které nebyly na základě kroku 1) určeny jako statisticky významné, mohou být významné v přítomnosti jiných proměnných. Proto sestavíme modely, které vzniknou z modelu M_n vždy přidáním jedné takové proměnné. Ty proměnné, u nichž došlo k významnému poklesu statistiky $-2ln\ell(\hat{\beta})$, budou do modelu M_n přidány. Vznikne model, který označíme jako M_m.

4) V posledním kroku se ujistíme, že již žádná z proměnných modelu M_m nemůže být odstraněna. Pokud, po odstranění nějaké proměnné, statisticky významně nenaroste hodnota $-2ln\ell(\hat{\beta})$, tuto proměnnou z modelu vyloučíme.

Pokud máme v souboru nějakou **primární vysvětlující proměnnou** (např. efekt léčebné metody, jestli je laparoskopická metoda lepší než otevřená apod.), budeme výše uvedenou strategií sestavovat model, u něhož nebudeme tuto proměnnou brát v potaz. Až dojdeme k modelu M_m, přidáme do modelu primární proměnnou.

Model

K sestavení modelu využijeme statistického softwaru ***SPSS Version 20***.

Než začneme model sestavovat musíme si uvědomit, že je rozdíl pracovat s **číselnou** proměnnou jako je např. *věk, BMI* apod. nebo pracovat s proměnnou, která je **kategoriální** jako např. *ASA, grading* apod. Přičemž pokud se jedná o dichotomickou proměnnou (např. *pohlaví: 0-muž, 1-žena*), můžeme na ni nahlížet stejně jako na číselné proměnné, avšak pokud máme více hodnot (např. *ASA: 1-4*), musíme každou úroveň překódovat pomocí nul a jedničky.

ASA pak nebude v modelu zastoupena jako jedna proměnná, ale rozpadne se na tři proměnné označené v tabulce 5 jako ASA_1, ASA_2, ASA_3. U pacienta, který má např. zaznamenanou hodnotu $ASA = 2$, dosadíme do modelu $ASA_1 = 1$, $ASA_2 = 0$, $ASA_3 = 0$. Podobně to bude i s jinými kategoriálními proměnnými. Více v [3].

úroveň		ASA_1	ASA_2	ASA_3
ASA	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

Tabulka 5: Kódování kategoriální proměnné

Musíme si uvědomit, že při sestavování modelu budeme např. za proměnnou ASA přidávat ve skutečnosti 3 proměnné, a tedy testování rozdílu statistik $-2ln\ell(\hat{\beta})$ nebude s jedním stupněm volnosti df, ale se třemi.

První krok

Sestavíme modely, které budou obsahovat vždy jednu proměnnou a budeme sledovat statistiku $-2ln\ell(\hat{\beta})$. Výsledky jsou uvedené v tabulce 6.

V tabulce se nachází hodnota nulového modelu $-2ln(\ell(0))$, s kterým porovnáváme nově sestavené modely. Na základě p-hodnoty vidíme, že na hladině významnosti $\alpha = 0.05$ jsou v tuto chvíli statisticky významné vysvětlující proměnné *stadium*, *ASA*, *věk*, *ICHS*, *grading*, *DM*, *BMI*, *délka operace*, *arytmie* a *hypertenze*.

Druhý krok

Sestavíme model, který bude obsahovat všechny proměnné, které jsme v prvním kroku identifikovali jako statisticky významné. Z něj sestavíme modely, které vzniknout odebráním jedné vysvětlující proměnné. Výsledky jsou uvedeny v tabulce 7.

Pro potřeby tabulky jsou použity zkratky *ST=stadium*, *GR=grading*, *DO=délka operace*, *AR=arytmie* a *HP=hypertenze*.

Z tabulky 7 vidíme, že se jako statisticky významné jeví vysvětlující proměnné *stadium*, *grading*, *věk* a *arytmie*. Proto si tyto proměnné ponecháme a ostatní z modelu vyloučíme.

Třetí krok

Protože jsme vyloučili některé proměnné. Musíme postup z druhého kroku zopakovat. Sestavíme model, který bude obsahovat proměnné *stadium*, *grading*, *věk* a *arytmie*. Z něj vždy vyloučíme jednu proměnnou a budeme sledovat nárůst statistiky $-2ln\ell(\hat{\beta})$. Výsledky jsou uvedeny v tabulce 8.

Vidíme, že nyní se jeví všechny proměnné statisticky významné, a proto je v modelu ponecháme.

Čtvrtý krok

Nyní budeme sestavovat modely, které vzniknou tak, že do modelu obsahující proměnné *stadium*, *věk*, *grading* a *arytmie* přidáme vždy jednu proměnnou, která se v prvním kroku zdála jako statisticky nevýznamná. Pokud se teď bude v přítomnosti těchto proměnných jevit statisticky významná, přidáme ji do modelu. Výsledky jsou uvedeny v tabulce 9.

Vidíme (Tab. 9), že v přítomnosti daných proměnných se na hladině významnosti $\alpha = 0.05$ nejeví žádná přidaná proměnná statisticky významně. Proto model ponecháme v původním tvaru a nebude nic přidávat.

Model	$-2ln\ell(\hat{\beta})$	**Pokles**	**df**	**p-hodnota**
nulový	5279.443			
stadium	5005.904	273.539	3	0
ASA	5256.695	22.748	3	< 0.001
věk	5262.062	17.381	1	< 0.001
ICHS	5262.593	16.850	1	< 0.001
DM	5268.932	10.511	1	0.001
grading	5266.918	12.525	2	0.002
BMI	5270.970	8.473	1	0.004
délka operace	5272.351	7.092	1	0.008
arytmie	5272.443	7.000	1	0.008
hypertenze	5275.212	4.231	1	0.040
cebrovaskulární	5275.896	3.547	1	0.060
krevní ztráta	5276.217	3.226	1	0.072
perop. komplikace	5277.006	2.437	1	0.119
pulmonální	5278.181	1.262	1	0.261
konverze	5276.963	2.480	2	0.289
renální	5278.341	1.102	1	0.294
předchozí operace	5278.779	0.664	1	0.415
pohlaví	5278.983	0.460	1	0.498
diagnóza	5278.439	1.004	2	0.605
jaterní	5279.334	0.109	1	0.741

Tabulka 6: Modely sestavené v prvním kroku

Model	$-2ln\ell(\hat{\beta})$	**Nárůst**	**df**	**p-hodnota**
ST+ASA+věk+ICHS+GR+DM+BMI+DO+AR+HP	4927.636			
-stadium (*ST*)	5217.911	290.275	3	0
-grading (*GR*)	4938.122	10.486	2	0.005
-věk	4934.398	6.762	1	0.009
-arytmie (*AR*)	4934.188	6.552	1	0.010
-BMI	4931.449	3.813	1	0.051
-délka operace (*DO*)	4930.506	2.870	1	0.090
-ASA	4933.654	6.018	3	0.111
-DM	4929.679	2.043	1	0.153
-hypertenze (*HP*)	4928.223	0.587	1	0.444
-ICHS	4927.846	0.210	1	0.647

Tabulka 7: Modely sestavené v druhém kroku

Pátý krok
V pátém kroku se ujistíme, že žádná proměnná v modelu, který obsahuje proměnné

Model	$-2ln\ell(\hat{\beta})$	**Nárůst**	**df**	**p-hodnota**
stadium+grading+věk+arytmie	4947.392			
-stadium	5245.556	298.164	3	0
-věk	4972.797	25.405	1	< 0.001
-arytmie	4958.492	11.100	1	0.001
-grading	4957.265	9.873	2	0.007

Tabulka 8: Modely sestavené ve třetím kroku

Model	$-2ln\ell(\hat{\beta})$	**Pokles**	**df**	**p-hodnota**
stadium+věk+grading+arytmie	4947.392			
krevní ztráta	4943.805	3.587	1	0.058
cebrovaskulární	4943.798	3.594	1	0.058
pulmonální	4945.087	2.305	1	0.129
perop. komplikace	4945.610	1.782	1	0.182
diagnóza	4944.176	3.216	2	0.200
jaterní	4945.854	1.538	1	0.215
konverze	4944.638	2.754	2	0.252
renální	4946.844	0.548	1	0.459
předchozí operace	4947.191	0.201	1	0.654
pohlaví	4947.304	0.088	1	0.767

Tabulka 9: Modely sestavené ve čtvrtém kroku

stadium, *věk*, *grading* a *arytmie*, se nedá vyloučit. Vzhledem k tomu, že se jedná o stejnou situaci jako v tabulce 8, žádná proměnná se již nedá vyloučit.

Šestý krok

V posledním kroku zahrneme do modelu primární proměnnou *skupina*, která určuje, zda byl pacient operován otevřeně či laparoskopicky. Výsledek je uveden v tabulce 10.

Vidíme, že vysvětlující proměnná *skupina* nevyšla statisticky významně. Dle modelu to, zda byl pacient operován otevřeně nebo laparoskopicky, nemá vliv na hazardní funkci daného pacienta. Nepotvrdila se nám tedy domněnka, že by laparoskopická operace byla pro pacienta statisticky lepší než operace otevřená.

Model	$-2ln\ell(\hat{\beta})$	**Pokles**	**df**	**p-hodnota**
stadium+věk+grading+arytmie	4947.392			
skupina	4945.889	1.503	1	0.220

Tabulka 10: Modely sestavené v šestém kroku

To, které proměnné budou zahrnuty do modelu, závisí značně na volbě strategie. Např. pokud bychom zvolili metodu **dopředního výběru**, pak by výsledný model zahrnoval proměnné *stadium*, *věk*, *arytmie*, *grading*, *ASA*, *krevní ztráta* a *délka operace*.

5.3 Závěry vyplývající z modelu

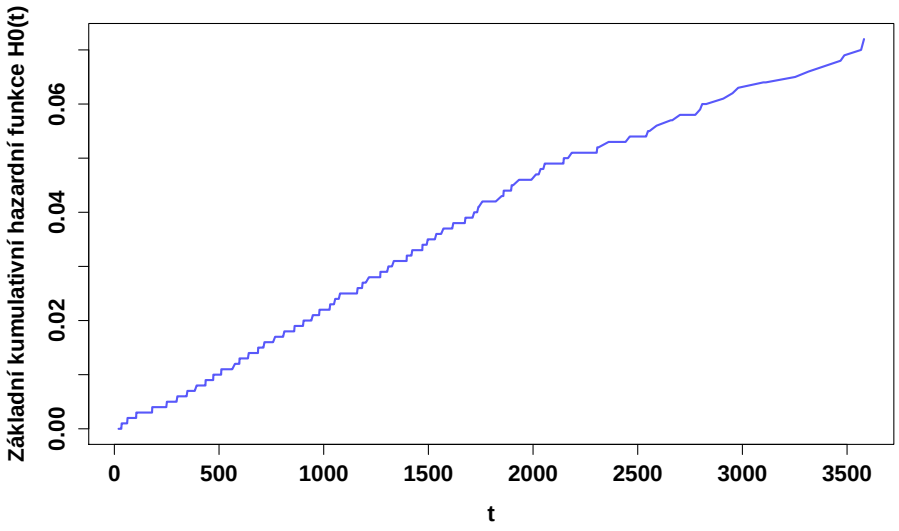
Pro model obsahující vysvětlující proměnné *stadium*, *arytmie*, *grading* a *věk* dostaneme následující odhady koeficientů β

	$\hat{\beta}$	$SE(\hat{\beta})$	$e^{\hat{\beta}}$	*95% CI pro* $e^{\hat{\beta}}$	
				dolní	*horní*
$stadium_1$	0.472	0.208	1.603	1.066	2.410
$stadium_2$	1.180	0.198	3.254	2.207	4.798
$stadium_3$	2.560	0.199	12.936	8.754	19.106
věk	0.026	0.005	1.026	1.016	1.036
arytmie	0.517	0.147	1.676	1.256	2.237
$grading_1$	-0.183	0.109	0.833	0.673	1.031
$grading_2$	0.278	0.159	1.321	0.969	1.801

Výsledný model má tvar

$$h(t) = h_0(t)e^{0.472ST_1+1.180ST_2+2.560ST_3+0.517AR+0.026vek-0.183GR_1+0.278GR_2},$$

kde ST reprezentuje $stadium$, AR $arytmii$ a GR $grading$.



Obrázek 4: Odhad základní kumulativní hazardní funkce $\hat{H}_0(t)$

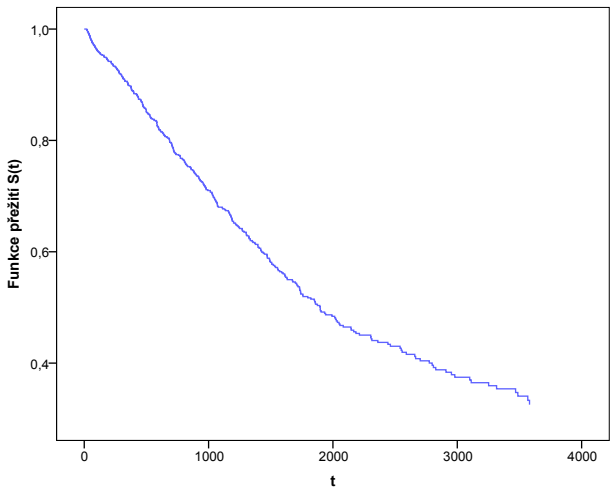
Obrázek 4 zachycuje odhad základní hazardní funce $\hat{h}_0(t)$. Povšimneme si, že všechny koeficienty $\hat{\beta}$ výjma koeficientu proměnné $grading_1$ jsou kladné. To znamená, že v přítomnosti odpovídajících proměnných se riziko pooperačních komplikací v porovnání se základní hazardní funkcí zvyšuje.

Hodnota $e^{\hat{\beta}}$ má také svoji interpretaci. Udává, kolikrát se zvýší riziko pooperačních komplikací oproti riziku daném základní hazardní funkcí, jestliže danou proměnnou navýšíme o jendotku. Jedná se vlastně o hazardní poměr. Např. pro $stadium_3$ v porovnání se základní hazardní funkcí (odpovídá pacientům z referenční skupiny - všechny naměřené hodnoty jsou rovny nule) dostaneme

$$\frac{h_1(t)}{h_0(t)} = \frac{h_0(t)e^{0.472\cdot 0+1.180\cdot 0+2.560\cdot 1+0.517\cdot 0+0.026\cdot 0-0.183\cdot 0+0.278\cdot 0}}{h_0(t)} = e^{2.560} = 12.936$$

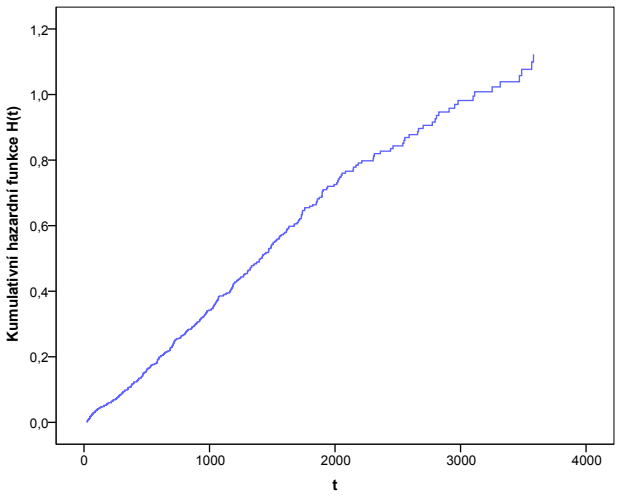
Statistický software **SPSS** umožňuje získat grafy funkce přežití a kumulativní hazardní funkce vyčíslené v hodnotách odpovídajícím středním hodnotám jednotlivých vysvětlujících proměnných. Neposkytuje sice přímo grafický výstup základní hazardní funkce, na druhou stranu aspoň umožňuje vyexportovat do excelu hodnoty základní kumulativní hazardní funkce $H_0(t)$. Jedná-li se o kategoriální proměnnou, lze vykreslit grafy funkce přežití a kumulativní hazardní funkce v závislosti na jednotlivých faktorech.

	$stadium_1$	$stadium_2$	$stadium_3$	*věk*	*arytmie*	$grading_1$	$grading_2$
Průměr	0.284	0.311	0.234	65.403	0.116	0.582	0.102



Obrázek 5: Funkce přežití vyčíslená v průměrech vysvětlujících proměnných

At'už se jedná o statistický software **SPSS** nebo **R**, oba vyčíslují nikoliv základní funkci přežití (resp. základní kumulativní hazardní funkci), ale vyčíslují funkci přežití (resp. kumulativní hazardní funkci) ve středních hodnotách (výběrový průměr) vysvětlujících proměnných. V našem případě obrázek 5 a 6.



Obrázek 6: Kumulativní haz. funkce vyčíslená v průměrech vysvětlujících proměnných

Vliv jednotlivých vysvětlujících proměnných

Když už sledujeme funkci přežití resp. kumulativní hazardní funkci ve středních hodnotách, je zajímavé zkoumat samostatný vliv jednotlivých vysvětlujících proměnných na tyto funkce.

Stadium

Na obrázcích 7 a 8 vidíme grafy funkce přežití a kumulativní hazardní funkce. Vysvětlující proměnná *stadium* odpovídá ohodnocení stadia onemocnění pacienta na stupnici $1-4$, kde 4 značí velmi vážné stadium. Výsledek vyplývající z modelu je očekávatelný. S horším stadiem se riziko pooperačních komplikací zvyšuje. Pacienti ve čtvrtém stádiu mají největší riziko pooperačních komplikací.

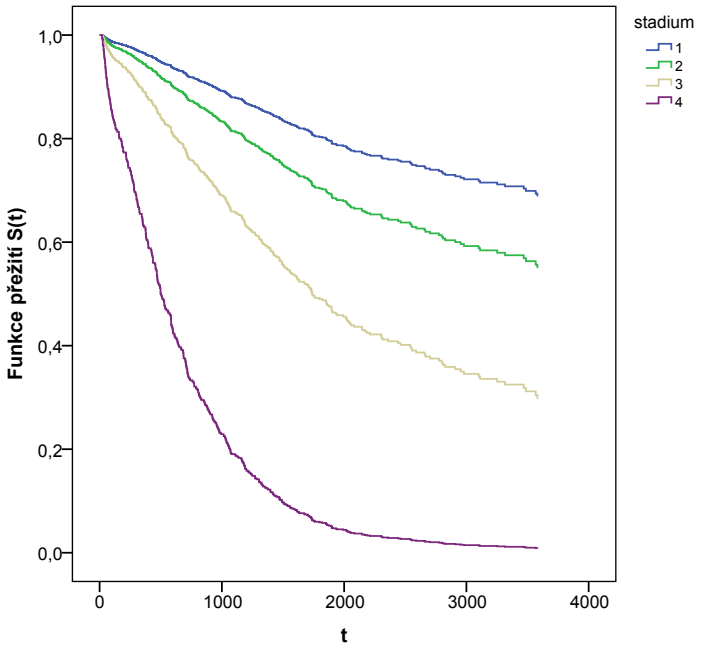
Pokud nás zájímá o kolik se riziko pooperačních komplikací u pacientů ve čtvrtém stádiu zvýší oproti pacientům v prvním stádiu, stačí spočítat hazardní poměr mezi hazardními fukcemi vzoru 4 a vzoru 1. Dostaneme pak $HR(t) = e^{2.56} = 12.936$. Tedy riziko pooperačních komplikací u pacienta ve čtvrtém stádiu bude téměř třináctkrát větší než u pacienta v prvním stádiu.

Arytmie

Na obrázku 9 a 10 vidíme grafy funkce přežití a kumulativní hazardní funkce. Z grafů či z tabulky s odhadnutými koeficienty $\hat{\beta}$ můžeme vyčíst, že u pacientů s arytmii je riziko pooperačních komplikací 1.676 krát větší než u pacientů bez arytmie.

		Vzor			
	Průměr	1	2	3	4
$stadium_1$	0.284	0.000	1.000	0.000	0.000
$stadium_2$	0.311	0.000	0.000	1.000	0.000
$stadium_3$	0.234	0.000	0.000	0.000	1.000
věk	65.403	65.403	65.403	65.403	65.403
arytmie	0.116	0.116	0.116	0.116	0.116
$grading_1$	0.582	0.582	0.582	0.582	0.582
$grading_2$	0.102	0.102	0.102	0.102	0.102

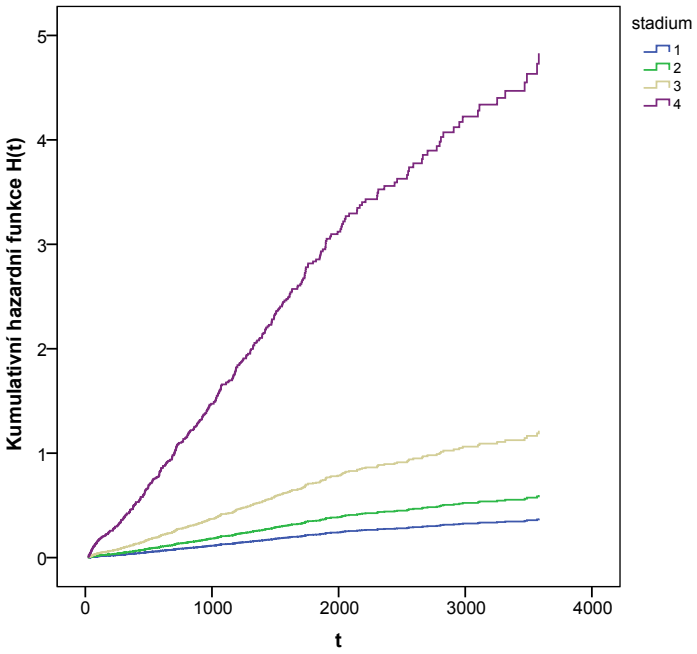
Tabulka 11: Hodnoty vstupních vysvětlujících proměnných - *Stadium*



Obrázek 7: Funkce přežití - *stadium*

Grading

Zde výsledky nesplňují naše očekávání. Vysvětlující proměnná *grading* totiž odpovídá ohodnocení míry rozšíření onemocnění. Čekali bychom, že se zvyšujícím se ohodnocením *grading* bude riziko pooperačních komplikací narůstat, avšak na základě modelu musíme konstatovat, že riziko pooperačních komplikací u pacientů s gradingem 1 je oproti pacientům s gradingem 0 menší a riziko pooperačních komplikací u pacientů s gradingem 2 je 1.321 krát větší než u pacientů s gradingem 0. Graficky tuto skutečnost zachycují obrázky 11 a 12.



Obrázek 8: Kumulativní hazardní funkce - *stadium*

		Vzor	
	Průměr	0	1
$stadium_1$	0.284	0.284	0.284
$stadium_2$	0.311	0.311	0.311
$stadium_3$	0.234	0.234	0.234
věk	65.403	65.403	65.403
arytmie	0.116	0	1
$grading_1$	0.582	0.582	0.582
$grading_2$	0.102	0.102	0.102

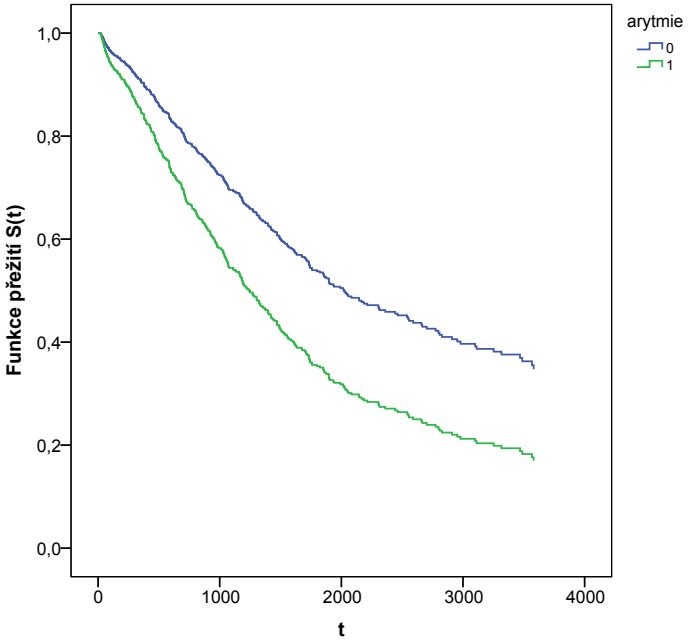
Tabulka 12: Hodnoty vstupních vysvětlujících proměnných - *Arytmie*

Věk

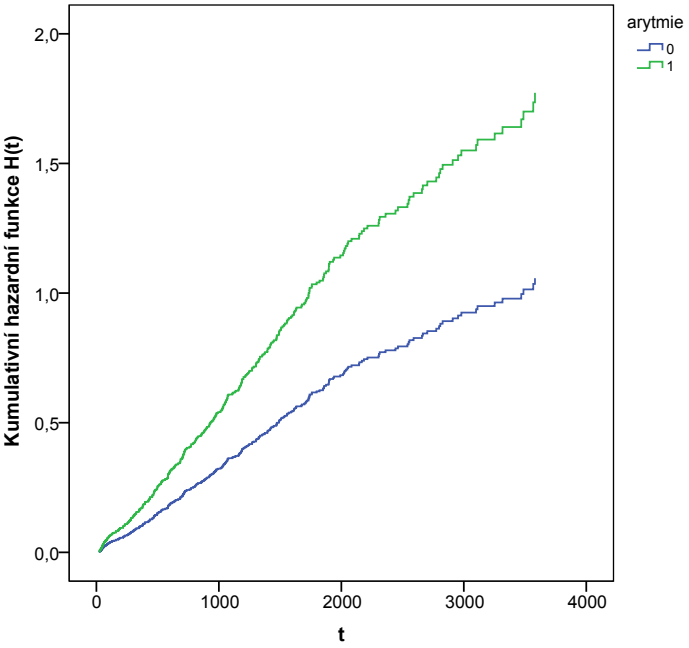
Poslední vysvtělující proměnnou, o které jsme zatím nemluvili, je *věk*. K této proměnné nebudeme ukazovat grafy funkce přežití ani kumulativní hazardní funkce, protože navýšíme-li danou proměnnou o jednotku, zvýší se riziko pooperačních komplikací 1.026 krát, což bude mít za následek, že takto dané grafy budou téměř totožné. Vidíme, že u pacienta o rok staršího nebude změna rizika pooperačních komplikací tak výrazná. Zajímavější bude sledovat, jak se změní riziko pooperačních komplikací u pacienta o 10 let staršího. V takovém případě dostaneme hazardní poměr

$$HR(t) = e^{0.026 \cdot 10} = 1.3$$

U pacienta o deset let staršího bude riziko pooperačních komplikací 1.3 krát větší. Věkový



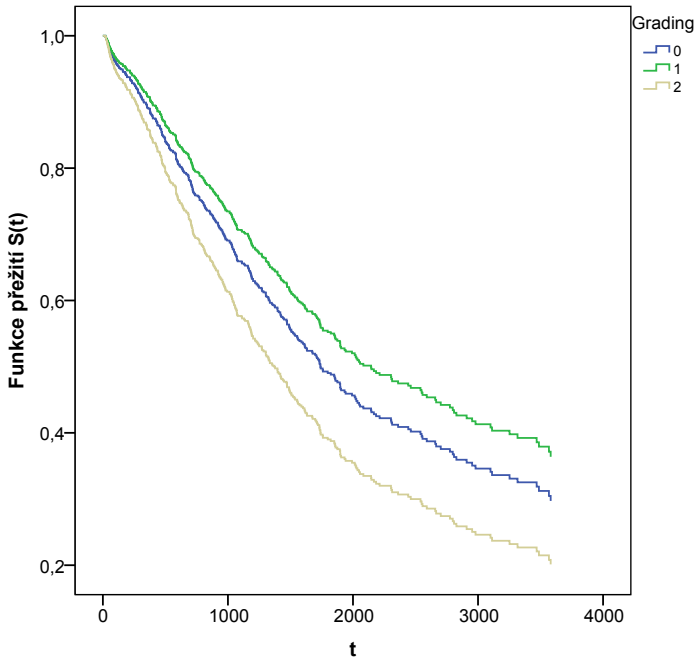
Obrázek 9: Funkce přežití - *arytmie*



Obrázek 10: Kumulativní hazardní funkce - *arytmie*

		Vzor		
	Průměr	0	1	2
$stadium_1$	0.284	0.284	0.284	0.284
$stadium_2$	0.311	0.311	0.311	0.311
$stadium_3$	0.234	0.234	0.234	0.234
věk	65.403	65.403	65.403	65.403
arytmie	0.116	0.116	0.116	0.116
$grading_1$	0.582	0.000	1.000	0.000
$grading_2$	0.102	0.000	0.000	1.000

Tabulka 13: Hodnoty vstupních vysvětlujících proměnných - *Grading*



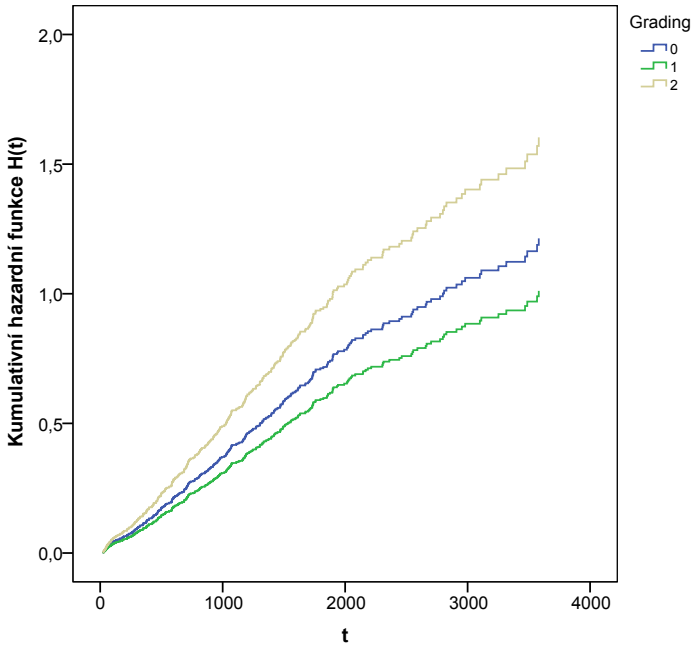
Obrázek 11: Funkce přežití - *grading*

rozdíl 30 let pak způsobí, že u takto staršího pacienta bude riziko poopearčních komplikací dokonce 2.18 krát větší.

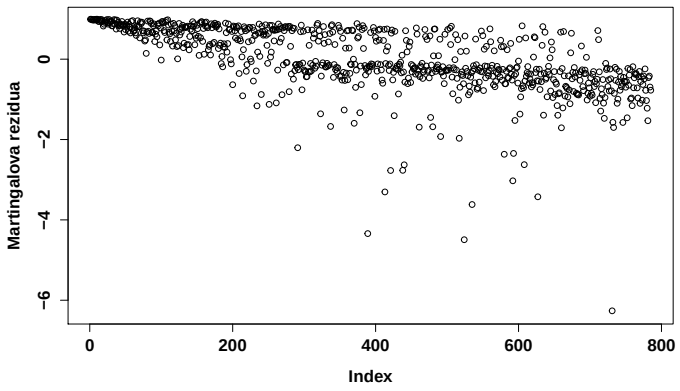
5.4 Ověření adekvátnosti modelu

Jekmile sestavíme model, musíme ověřit, jestli je tento model adekvátní. Existuje řada metod, které realizují postup ověřování. Ty nejpoužívanější jsou založeny na **reziduích**.

Martingalova rezidua zvýrazňují ty jedince, jejichž časy přežití nebyly modelem příliš dobře zachyceny. Pro náš datový soubor, kde pacienti byli seřazeni podle času pooperačních komplikací, jsou martingalova rezidua graficky znázorněna na obrázku 13.



Obrázek 12: Kumulativní hazardní funkce - *grading*

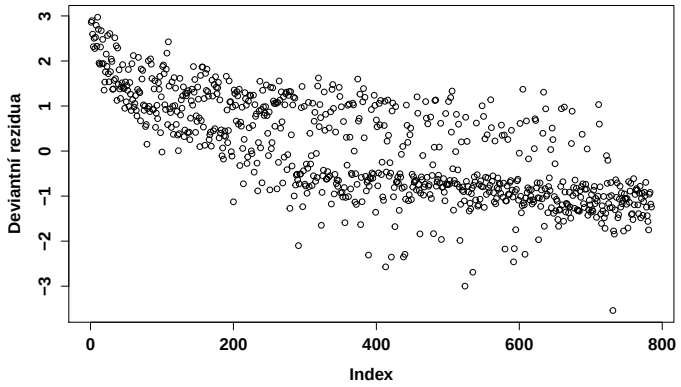


Obrázek 13: Martingalova rezidua

Jestliže u pacienta je vypočtena velká záporná hodnota rezidua, znamená to, že u něj došlo k pooperačním komplikacím daleko později, než bychom mu na základě modelu predikovali. Naopak pokud je reziduum blízké jedničce, znamená to, že u pacienta došlo k pooperačním komplikacím podstatně dříve, než bychom na základě modelu určili. V případě, že absolutní hodnota rezidua je neobvykle velká, bude muset být příslušný subjekt podroben další analýze.

Z obrázku 13 vidíme, že si náš model nevede špatně, ale přesto je tu dle mého názoru

prostor pro vylepšení. Hodnoty, u nichž bychom predikovali kratší dobu do pooperačních komplikací než ve skutečnosti byla, nejsou tak časté, ale naopak je tu značné množství hodnot blízké jedničce, které zachycují pacienty u nichž bychom predikovali delší dobu do pooperačních komplikací než ve skutečnosti byla.



Obrázek 14: Deviantní rezidua

Podobnou interpretaci jako předchozí rezidua mají i **rezidua deviantní**, která jsou graficky znázorněna na obrázku 14. Autoři často označují tento typ reziduí za více symetrická kolem nuly (předchozí rezidua nemusejí být symetrická kolem nuly ani v případě, že je model adekvátní), a tedy vhodnější k posouzení adekvátnosti modelu.

V našem případě musíme bohužel konstatovat, že na základě deviantních reziduí, bychom měli hledat adekvátnější model, protože deviance D, která je dána jako součet druhých mocnin deviantních reziduí, je příliš velká ($D = 1031.422$), což naznačuje nevhodnost modelu. Více k teorii reziduí lze nalézt v [3], [10] nebo [14].

5.5 Program

Cílem práce je také vytvořit funkční program, který sestaví Coxův proporcionální hazardní model a vygeneruje vyhodnocení modelu na základě příslušných dat. Pro program byl zvolen statistický software **R** a prostředí **RStudia**.

R version 3.0.2
Platform 32-bit
R studion version 0.97.551

Aby byla zaručena funkčnost programu, je třeba mít nainstalované následující doplňující knihovny:

- **survival** - knihovna s metodami pro analýzu přežití (Kaplna-Meierův odhad, Coxův proporcionální hazardní model)

- **R2HTML** - knihovna podporující generování výstupu do formátu html
- **stringr** - knihovna usnadňující práci s řetězci

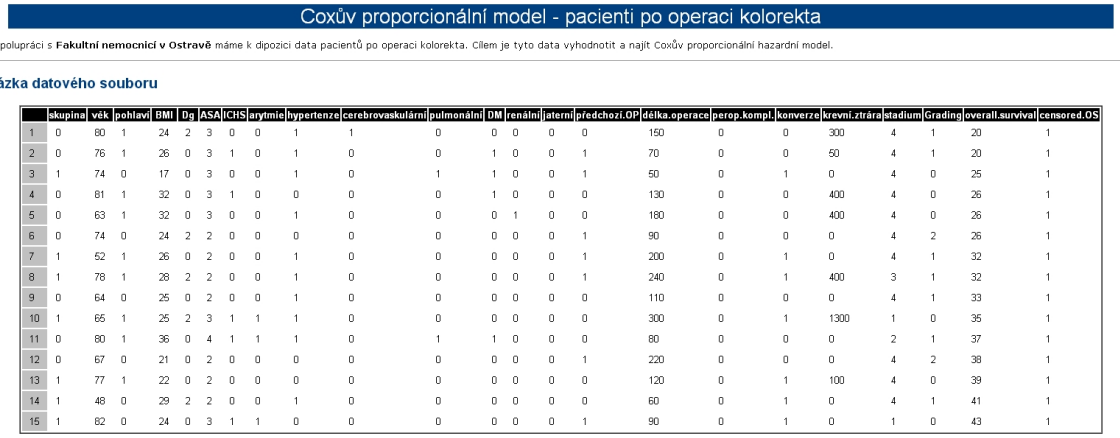
Další použité knihovny: *datasets*, *graphics*, *grDevices*, *methods*, *stats* a *utils*.

V programu je použita **Efronova aproximace** částečné věrohodnostní funkce a **metoda dopředného výběru** vysvětlujících proměnných.

Data vstupující do programu musí být ve formátu .csv, který používá středník jako oddělovač. Daný soubor může obsahovat pouze jeden list, který ukládá odpovídající data. Ostatní nadbytečné listy/záložky musí být ze souboru odstraněny. Předpokládá se, že vstupující soubor zachovává původní názvy datových sloupců. Přestože původní soubor obsahuje daleko více proměnných než je uvedeno v seznamu v kapitole 5.1, program pracuje pouze s těmito vybranými vysvětlujícími proměnnými.

Hlavním výstupem je soubor *PH model - vyhodnocení_main.html*, který zahrnuje základní popis dat, Coxův proporcionální hazardní model (odhady koeficientů, hazardní poměry), odhad základní kumulativní hazardní funkce a základní funkce přežití (grafy), vliv jednotlivých vysvětlujících proměnných na základní kumulativní hazardní funkci resp. základní funkci přežití, grafické porovnání základní funkce přežití a funkce přežití získanou Kaplan-Meierovým odhadem bez vlivu vysvětlujících proměnných.

Ukázka výstupu z programu je na obrázku 15. Další lze nalézt v příloze A.



Coxův proporcionální model - pacienti po operaci kolorekta

Ve spolupráci s **Fakultní nemocnicí v Ostravě** máme k dipozici data pacientů po operaci kolorekta. Cílem je tyto data vyhodnotit a najít Coxův proporcionální hazardní model.

Ukázka datového souboru

	skupina	věk	pohlaví	BMI	Dg	ASA	ICHS	arytmie	hypertenze	cerebrovaskulární	pulmonální	DM	renální	jaterní	předchozí.OP	délka.operace	perop.kompl.	konverze	krevní.ztráta	stadium	Grading	overall.survival	censored.OS
1	0	80	1	24	2	3	0	0	1	1	0	0	0	0	0	150	0	0	300	4	1	20	1
2	0	76	1	26	0	3	1	0	1	0	0	1	0	0	1	70	0	0	50	4	1	20	1
3	1	74	0	17	0	3	0	0	1	0	1	1	0	0	1	50	0	1	0	4	0	25	1
4	0	81	1	32	0	3	1	0	0	0	0	1	0	0	0	130	0	0	400	4	0	26	1
5	0	63	1	32	0	3	0	0	1	0	0	0	1	0	0	180	0	0	400	4	0	26	1
6	0	74	0	24	2	2	0	0	0	0	0	0	0	0	1	90	0	0	0	4	2	26	1
7	1	52	1	26	0	2	0	0	1	0	0	0	0	0	1	200	0	1	0	4	1	32	1
8	1	78	1	28	2	2	0	0	1	0	0	0	0	0	1	240	0	1	400	3	1	32	1
9	0	64	0	25	0	2	0	0	1	0	0	0	0	0	0	110	0	0	0	4	1	33	1
10	1	65	1	25	2	3	1	1	1	0	0	0	0	0	0	300	0	1	1300	1	0	35	1
11	0	80	1	36	0	4	1	1	1	0	1	1	0	0	0	80	0	0	0	2	1	37	1
12	0	67	0	21	0	2	0	0	0	0	0	0	0	0	1	220	0	0	0	4	2	38	1
13	1	77	1	22	0	2	0	0	0	0	0	0	0	0	0	120	0	1	100	4	0	39	1
14	1	48	0	29	2	2	0	0	1	0	0	0	0	0	0	60	0	1	0	4	1	41	1
15	1	82	0	24	0	3	1	1	0	0	0	0	0	0	1	90	0	1	0	1	0	43	1

Obrázek 15: Ukázka výstupu z programu

6 Závěr

V rámci práce jsme se seznámili s nejpoužívanějšími metodami v analýze přežití. Ukázali jsme si různé parametrické i neparametrické modely a vysvětlili jsme si rozdíl mezi AFT a PH modely. Následně jsme pozornost věnovali Coxovu proporcionálnímu hazardnímu modelu, který řadíme mezi PH modely. Odvodili jsme částečnou věrohodnostní funkci a ukázali jsme si její aproximace v případě opakovaných pozorovaných časů událostí (Breslowova aproximace, Efronova aproximace). Dále byla představena MLE metoda odhadu koeficientů v částečné věrohodnostní funkci, metody pro vyhodnocení významnosti odhadnutých koeficientů v modelu (test poměru částečné věrohodnostní funkce, Waldova statistika, Skóre test) a v neposlední řadě, přestože běžně nepotřebujeme znát základní hazardní funkci, byl ukázán i její odhad.

V praktické části jsme pak měli možnost nastínit strategie výběru vhodných vysvětlujících proměnných do modelu a vyhodnotit poskytnutá data. Ukázalo se, že ne všechny proměnné mají statisticky významný vliv na hazardní funkci pacienta. Jako nejvýznamnější se ukázaly vysvětlující proměnné *stadium*, *arytmie*, *grading* a *věk*. Přičemž nejvíce hazardní funkci pacienta ovlivňuje proměnná *stadium* a to tak, že může riziko pooperačních komplikací navýšit u daného pacienta až třináctkrát. Předpoklad, že by laparoskopická metoda měla být pro pacienta méně rizikovější než metoda otevřená, se nám nepotvrdil. Na základě dat musíme konstatovat, že proměnná *skupina* nemá statisticky významný vliv na změnu hazardní funkce pacienta. Avšak je tu stále prostor pro nalezení adekvátnějšího modelu. Toho bychom mohli docílit např. změnou strategie výběru vhodných vysvětlujících proměnných do modelu.

Vytvořili jsme také program ve statistickém softwaru R, který dokáže sestavit Coxův proporcionální hazardní model pomocí metody dopředného výběru na základě testu poměru částečné věrohodnostní funkce. Program uživateli vygeneruje *html* soubor s vyhodnocením poskytnutých dat. V programu je zahrnutý základní popis dat, výstupní tabulka modelu (odhady koeficientů, jejich chyby, hazardní poměry) a grafy vlivu jednotlivých vysvětlujících proměnných na funkci přežití a kumulativní hazardní funkci.

Téma Coxova proporcionálního hazardního modelu nabízí stále prostor pro další studium. Do budoucna bych se chtěla zaměřit na postupy, které umožní sestavit adekvátnější model než ten, který byl odvozen v rámci této prace. Problém nalezení nejlepšího modelu není triviální. Tato dovednost se velmi cení a je naprosto běžné, že se statistici předhánějí v tom, kdo přijde s modelem, který nejlépe zachycuje danou situaci a dokáže poměrně velmi přesně predikovat.

Žaneta Miklová

7 Reference

[1] BIAN, Hui. *Survival Analysis Using SPSS*, [online]. 2013 [cit. 2014-04-08]. Dostupné z: http://core.ecu.edu/ofe/StatisticsResearch/Survival Analysis Using SPSS.pdf

[2] BRIŠ, Radim a Martina LITSCHMANNOVÁ. *Statistika II.* Ostrava: Vysoká škola báňská - Technická univerzita, 2007, 1 CD-R. ISBN 978-80-248-1482-7.

[3] COLLETT, D. *Modelling survival data in medical research.* 2nd ed. Boca Raton, Fla.: Chapman, c2003, 391 p. Texts in statistical science. ISBN 15-848-8325-1.

[4] DIEZ, David. *Survival Analysis in R*, [online]. 2007 [cit. 2014-04-08]. Dostupné z: http://anson.ucdavis.edu/ hiwang/teaching/10fall/R_tutorial 1.pdf

[5] FABSIC, Peter, Vakhrushev EVGENY a Kevin ZEMMER. *The Cox Proportional Hazard Model and Its Characteristics.*, [online]. 2011 [cit. 2014-04-08]. Dostupné z: http://stat.ethz.ch/education/semesters/ss2011/seminar/contents/handout_3.pdf nebo https://stat.ethz.ch/education/semesters/ss2011/seminar/contents/presentation_3.pdf

[6] FÜRSTOVÁ, Jana. *Metody analýzy přežití.* [online], 2010. [cit. 2014-04-08]. Dostupné z: http://www.ejbi.org/en/ejbi/article/21-cs-metody-analyzy-preziti.html

[7] HARCEK, Martin. *Neparametrické a semiparametrické metódy odhadu Value at Risk.*, Bratislava, 2010 [cit.2014-03-06]. Diplomová práce. Univerzita Komenského v Bratislavě, Fakulta matematiky, fyziky a informatiky. Vedoucí práce doc. Mgr. Marián Grendár, PhD.

[8] HOLUBOVÁ, Kamila. *Modelování rizikové funkce v populačním hodnocení přežití.*, Brno, 2011 [cit.2014-03-06]. Diplomová práce. Masarykova univerzita, Přírodovědecká fakulta. Vedoucí práce doc. RNDr. Tomáš Pavlík, Ph.D.

[9] HONGSHIK, Ahn. *Log-normal Regression Modeling through Recursive Partitioning.*, [online]. Jefferson, Arkansas, U.S.A., 2011 [cit. 2014-04-08]. Dostupné z: http://www.ams.sunysb.edu/h̃ahn/psfile/paplogno.pdf

[10] HOSMER, David W, Stanley LEMESHOW a Susanne MAY. *Applied survival analysis: regression modeling of time-to-event data.* 2nd ed. Hoboken, N.J.: Wiley-Interscience, c2008, xiii, 392 p. ISBN 04-717-5499-4.

[11] CHABIČOVSKÝ, Martin. *Statistická analýza rozdělení extrémních hodnot pro cenzorovaná data.*, [online]. Brno, 2011 [cit.2014-03-06]. Dostupné z: https://dspace.vutbr.cz/bitstream/handle/11012/4350/diplomovaprace.pdf. Diplomová práce. Vysoké učení technické v Brně, Fakulta strojního inženýrství. Vedoucí práce doc. RNDr. Jaroslav Michálek, CSc.

[12] KALBFLEISCH, J a Ross L PRENTICE. *The statistical analysis of failure time data.*, New York: Wiley, c1980. ISBN 04-710-5519-0.

[13] KOHOUT, Václav. *Teorie odhadu: Skriptum ZCU*, [online]. ZČU Plzeň, 2004 [cit. 2014-03-09]. Dostupné z: http://www.kmt.zcu.cz/person/Kohout/info_soubory/letnisem/zs/stat10.pdf

[14] LI, Yi. *Model Selection in Survival Analysis*, [online]. 2012 [cit. 2014-04-08]. Dostupné z: http://www-personal.umich.edu/ỹili/lect6notes.pdf

[15] LITSCHMANNOVÁ, Martina. *Úvod do statistiky*, [online]. Ostrava, 2011 [cit. 2014-03-09]. Dostupné z: http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/uvod_do_statistiky.pdf

[16] LITSCHMANNOVÁ, Martina. *Vybrané kapitoly z pravděpodobnosti*, [online]. Ostrava, 2011 [cit. 2014-03-09]. Dostupné z: http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/vybrane_kapitoly_pravdepodobnost.pdf

[17] NGUYEN, Vinh. *Regression Diagnostics for the Proportional Hazards model*, [online]. 2012 [cit. 2014-04-08]. Dostupné z: http://www.ics.uci.edu/ṽqnguyen/stat255/Lecture10.pdf

[18] NORUŠIS, Marija J. *SPSS statistics 17.0 guide to data analysis.* Upper Saddle River, N.J: Prentice Hall, 2008. ISBN 978-032-1621-436.

[19] RAO, C. Radhakrishna, Helge TOUTENBURG, SHALABH, Christian HEUMANN a M. SCHOMAKER. *Linear models and generalizations: least squares and alternatives.* 3rd extended ed. New York: Springer, c2008, xix, 570 p. ISBN 978-354-0742-265.

[20] REISNEROVÁ, Soňa. *Analýza přežití a Coxův model pro diskrétní čas*, [online]. 2004 [cit. 2014-04-08]. Dostupné z: http://www.statspol.cz/oldstat/robust/robust2004/reisnerova.pdf

[21] SELINGEROVÁ. *Neparametrické odhady podmíněné rizikové funkce* [online]. Ústav matematiky a statistiky, Přírodovědecká fakulta Masarykova univerzita, 2013 [cit. 2014-03-06]. Dostupné z: https://www.math.muni.cz/ amathnet/presentations/Selingerova.pdf

[22] SVEINBJORNSSON, Gardar, Jongkil KIM a Yongsheng WANG. *The Cox model in R*, [online]. 2011 [cit. 2014-04-08]. Dostupné z: http://stat.ethz.ch/education/semesters/ss2011/seminar/contents/handout_7.pdf

[23] TEETOR, Paul. *R cookbook.*, Sebastopol: O'Reilly Media, 2011, xviii, 413 s. ISBN 978-0-596-80915-7.

[24] UHER, Michal. *Parametrické modely v analýze přežití.*, [online]. Brno, 2011 [cit.2014-03-06]. Dostupné z: http://is.muni.cz/th/323663/prif_b/BP.pdf. Bakalářská práce. Masarykova Univerzita, Přírodovědecká fakulta. Vedoucí práce RNDr. Tomáš Pavlík, Ph.D.

[25] VAŠÍČKOVÁ, Hana. *Analýza přežití pro aktuální onkologická data.*, [online]. Ostrava, 2013 [cit.2014-03-06]. Dostupné z: http://www.fei.vsb.cz/export/sites/fei/k470/cs/theses/mgr/2013/VAS286.pdf. Diplomová práce. VŠB - Technická univerzita Ostrava, Fakulta elektrotechniky a informatiky. Vedoucí práce prof. Ing. Radim Briš, CSc.

[26] VOLF, Petr. *Regresní modely v analýze přežití.*, [online]. Praha, 1992 [cit.2014-03-06]. Dostupné z: http://www.statspol.cz/robust/1992_volf__92.pdf. ÚTIA ČSAV, Praha

[27] ZHANG, Daowen. *Analysis of Survival Data* [online]. North Carolina State University, Department of Statistics, 2005 [cit. 2014-03-20]. Dostupné z: http://www4.stat.ncsu.edu/ dzhang2/st745/chap7.pdf

A Ukázky výstupu z programu

Coxův proporcionální hazardní model

Pro Coxův proporcionální hazardní model jsme využili Efronovu aproximaci částečné věrohodnostní funkce a metodu dopředného výběru k určení významných vysvětlujících proměnných.

Jako statisticky významné byly určeny tyto proměnné:

1) stadium

2) vek

3) arytmie

4) grading

5) ASA

6) krevni.ztrata

7) delka.OP

Neprokázalo se, že by na hazardní funkci pacienta měl vliv druh operace (laparoskopická vs. otevřená).

	koeficienty
stadium_1	0.4362
stadium_2	1.1538
stadium_3	2.5547
vek_1	0.0194
arytmie_1	0.3659
grading_1	-0.1636
grading_2	0.3439
ASA_1	0.1080
ASA_2	0.4455
ASA_3	0.3672
krevni.ztrata_1	0.0004
delka.OP_1	-0.0022

Obrázek 16: Příloha - ukázka výstupu z programu 1

Coxův proporcionální model - pacienti po operaci kolorekta

Ve spolupráci s **Fakultní nemocnicí v Ostravě** máme k dipozici data pacientů po operaci kolorekta. Cílem je tyto data vyhodnotit a najít Coxův proporcionální hazardní model.

Ukázka datového souboru

	skupina	věk	pohlaví	BMI	Dg	ASA	ICHS	arytmie	hypertenze	cerebrovaskulární	pulmonální	DM	renální	jaterní	předchozí.OP	délka.operace	perop.kompl.	konverze	krevní.ztráta	stadium	Grading	overall.survival	censored.OS
1	0	80	1	24	2	3	0	0	1	1	0	0	0	0	0	150	0	0	300	4	1	20	1
2	0	76	1	26	0	3	1	0	1	0	0	1	0	0	1	70	0	0	50	4	1	20	1
3	1	74	0	17	0	3	0	0	1	0	1	1	0	0	1	50	0	1	0	4	0	25	1
4	0	81	1	32	0	3	1	0	0	0	0	1	0	0	0	130	0	0	400	4	0	26	1
5	0	63	1	32	0	3	0	0	1	0	0	0	1	0	0	180	0	0	400	4	0	26	1
6	0	74	0	24	2	2	0	0	0	0	0	0	0	0	1	90	0	0	0	4	2	26	1
7	1	52	1	26	0	2	0	0	1	0	0	0	0	0	1	200	0	1	0	4	1	32	1
8	1	78	1	28	2	2	0	0	1	0	0	0	0	0	1	240	0	1	400	3	1	32	1
9	0	64	0	25	0	2	0	0	1	0	0	0	0	0	0	110	0	0	0	4	1	33	1
10	1	65	1	25	2	3	1	1	1	0	0	0	0	0	0	300	0	1	1300	1	0	35	1
11	0	80	1	36	0	4	1	1	1	0	1	1	0	0	0	80	0	0	0	2	1	37	1
12	0	67	0	21	0	2	0	0	0	0	0	0	0	0	1	220	0	0	0	4	2	38	1
13	1	77	1	22	0	2	0	0	0	0	0	0	0	0	0	120	0	1	100	4	0	39	1
14	1	48	0	29	2	2	0	0	1	0	0	0	0	0	0	60	0	1	0	4	1	41	1
15	1	82	0	24	0	3	1	1	0	0	0	0	0	0	1	90	0	1	0	1	0	43	1

Obrázek 17: Příloha - ukázka výstupu z programu 2

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
stadium_1	4.4e-01	1.5e+00	2.1e-01	2.1e+00	3.6e-02
stadium_2	1.2e+00	3.2e+00	2.0e-01	5.8e+00	6.1e-09
stadium_3	2.6e+00	1.3e+01	2.0e-01	1.3e+01	0.0e+00
vek_1	1.9e-02	1.0e+00	5.8e-03	3.3e+00	8.2e-04
arytmie_1	3.7e-01	1.4e+00	1.6e-01	2.3e+00	1.9e-02
grading_1	-1.6e-01	8.5e-01	1.1e-01	-1.5e+00	1.4e-01
grading_2	3.4e-01	1.4e+00	1.6e-01	2.1e+00	3.3e-02
ASA_1	1.1e-01	1.1e+00	2.0e-01	5.3e-01	6.0e-01
ASA_2	4.5e-01	1.6e+00	2.2e-01	2.0e+00	4.2e-02
ASA_3	3.7e-01	1.4e+00	3.5e-01	1.1e+00	2.9e-01
krevni.ztrata_1	4.0e-04	1.0e+00	1.3e-04	3.0e+00	3.0e-03
delka.OP_1	-2.2e-03	1.0e+00	8.1e-04	-2.7e+00	7.2e-03

V tabulce vidíme v prvním sloupečku odhady koeficientů β. Druhý sloupeček odpovídá hazardnímu poměru HR hazardních funkcí h1(t) a h0(t), kde v h1(t) jsou vstupní vysvětlující proměnné všechny rovny nule kromě příslušné proměnné (na řádku), která se rovná jedné. A h0(t) odpovídá základní hazardní funkci (všechny vstupní vysvětlující proměnné jsou rovny nule).

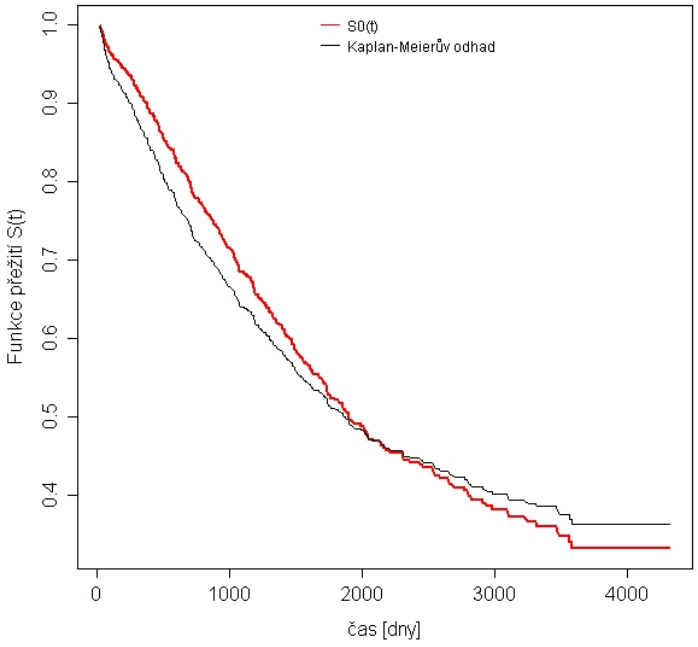
Všechny funkce jsou vyčíslené ve středních hodnotách (výběrový průměr) daných vysvětlujících proměnných. Nejedná se tedy o základní funkci přežití resp. základní kumulativní hazardní funkci!

	prumery
stadium_1	0.284
stadium_2	0.311
stadium_3	0.234
vek_1	65.403
arytmie_1	0.116
grading_1	0.582
grading_2	0.102
ASA_1	0.494
ASA_2	0.366
ASA_3	0.033
krevni.ztrata_1	218.981
delka.OP_1	161.283

Obrázek 18: Příloha - ukázka výstupu z programu 3

Funkce přežití ve středních hodnotách

Funkce přežití ve středních hodnotách odpovídá funkci přežití pro pacienty s naměřenými hodnotami danými tabulkou výše. V grafu je pro porovnání zahrnuta i křivka odhadu Kaplan-Meierovou metodou bez vlivu vysvětlujících proměnných.

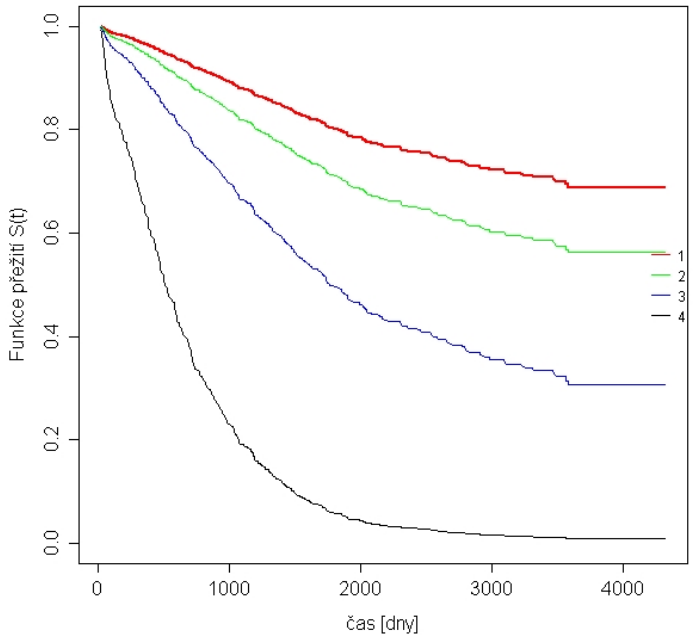


Obrázek 19: Příloha - ukázka výstupu z programu 4

Funkce přežití u jednotlivých proměnných

Pro každou vysvětlující proměnnou, která se ukázala jako významná vykreslíme grafy příslušných funkcí přežití. Jedná-li se o faktor, zahrneme odpovídající proměnné do jednoho grafu. Grafy zachycují, jak se změní funkce přežití ve středních hodnotách v závislosti na změně dané vysvětlující proměnné.

Funkce přežití: stadium

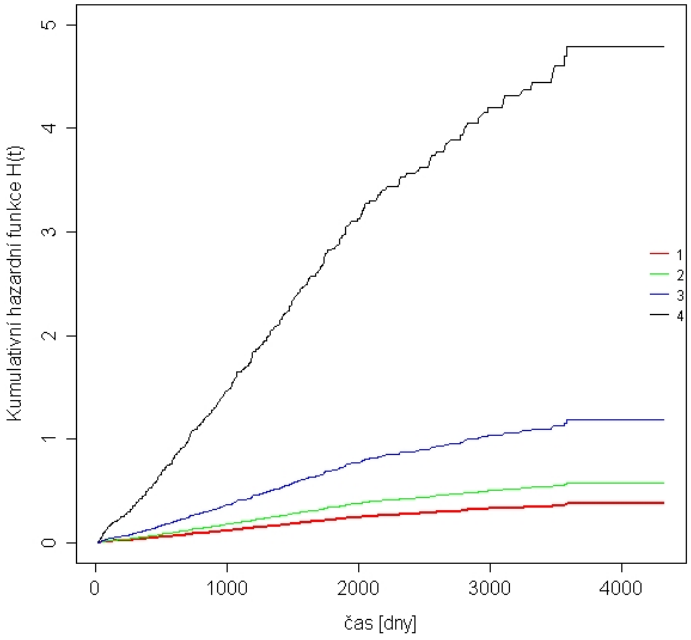


Obrázek 20: Příloha - ukázka výstupu z programu 5

Kumulativní hazardní funkce u jednotlivých proměnných

Grafy zachycují, jak se změní kumulativní hazardní funkce v závisloti na změně dané vysvětlující proměnné.

Kumulativní hazardní funkce: stadium



Obrázek 21: Příloha - ukázka výstupu z programu 6